52-82

333522

N91-70700

p 73

NASA KSC

Intelligent Interactive High Speed Data Search System

11 - 28 - 203

Annual Report
1990

Principal Investigator

Dr. James R. Driscoll

Associate Professor
College of Arts & Sciences
Department of Computer Science
University of Central Florida

(407) 823-2341

TABLE OF CONTENTS

Section 1.

Overview

The overall objective of this three year Cooperative Agreement project is to research a system capable of high speed search of very large text databases in order to (1) produce answers to questions expressed in natural language, and (2) generate summaries of database information. The objectives listed for Year 1 of this project are to essentially prepare for item (1) above. A somewhat detailed, but readable, three page description of item (1) appears in Section 2 of this report. During Year 1, this three page description was submitted to NASA Headquarters as a KSC project white paper for the Advanced Operations Program managed by Chuck Holliman.

The objectives for Year 1 have been achieved. However, there has been a crucial modification. Very simply, and with slight rewording from the Cooperative Agreement, the Year 1 objectives have been the following:

a. Design an internal format for analyzed text.

b. Develop a prototype text parsing system in regard to the format.

c. Establish the "dictionary" required for the prototype text parsing system.

d. Study the structure of English questions.

e. Purchase required PC platforms.

f. Put in place a high speed basic (shell) text retrieval system in preparation for extension to answering questions.

The crucial modification mentioned earlier affects objectives (a), (b), and (c). In order to extend our work beyond the performance of a prototype and to make our work applicable (without modification) to all text, we have eliminated the need (for now) to do precise parsing of English sentences and questions. Instead, our modification is to make use of basic, general world knowledge and improve well-known text retrieval techniques to accomplish our objective of being able to generate an answer to an English question.

A description of the modification affecting objectives (a), (b), and (c) can be found in Section 3 of this report. We are just now beginning to implement the procedures identified there. Preliminary experiments have shown good results. This implementation project will yield a Master's degree for Edgar Wendlandt.

Section 4 of this report is aimed at objective (c). Described there is a proposal which has recently been funded by the State of Florida High Technology and Industry Council (FHTIC). The intent of the funding is to develop a "dictionary" of general world knowledge and carry our research beyond the prototype stage and into commercialization. This funding begins in January 1991. The proposal was submitted in August of 1990 by Dr. Driscoll.

Section 5 presents an analysis of English question structure in regard to the general world knowledge aspect of our research. The report is somewhat technical and really not intended for a general audience but is included here solely to provide evidence of work on objective (d) of Year 1. This report was essentially generated by undergraduate student Gloria Corbett as an independent study project. No NASA funds were used to support this activity.

In regard to objectives (e) and (f), we have purchased two 33 MHz 386 PC clones, each with a 165 Mbyte, 14.5 ms hard drive. We have developed a high speed basic text retrieval system, and recently installed the system on one of these two machines. The retrieval system was implemented in the C programming language by three undergraduate students: Frank Fenneran, Bill Gellerstedt, and Mike Cuccarese. This activity was done as a classroom project and under several independent study agreements. No NASA funds were used to support this activity. The system is modeled after the French commercial product SPIRIT which is described in Section 6 of this report where a recent publication by Dr. Driscoll is provided.

Finally, to conclude this overview of project progress during Year 1 of the Cooperative Agreement, the following is a list of presentations during 1990 where progress reports were submitted to KSC.

Purpose	Date	Location
1. Briefing	20 February 1990	UCF
2. Project Review	20 March 1990	UCF
3. Video-Teleconference	23-24 April 1990	NASA Headquarters, JSC, KSC
4. Project Review	20 July 1990	KSC
5. Project Review	19 October 1990	UCF

Section 2.

White Paper Description of Project submitted to NASA Headquarters for the Advanced Operations Program

INTELLIGENT INTERACTIVE HIGH SPEED DATA SEARCH SYSTEM

KSC

James R. Driscoll

Department of Computer Science
University of Central Florida
P.O. Box 25000, Orlando, FL 32816
(407) 275-2341

This is a new project. It concerns natural language access to large amounts of textual information. The objective is to automate the access of textual information so that users are not overwhelmed by the amount of information searched. Currently, the project is motivated by the desire to provide convenient access to information contained in the numerous and large public information documents maintained by Public Affairs at KSC.

The documents maintained by Public Affairs at KSC consist of press releases and other printed information created at KSC and other NASA offices using various word processors. There are also documents from outside contractors, such as Rockwell, which produces the "NASA National Space Transportation System Reference" more often called the "shuttle manual". During a launch at KSC, about a dozen NASA employees access these printed documents to answer media questions. The amount of text is visually overwhelming. The shuttle manual alone is just over 1,000 pages (it is three inches thick). The planned document storage for NASA KSC Public Affairs is around 300,000 pages (approximately seventy-five feet of stacked pages) with about 5,000 pages added or replaced annually.

In regard to the operation of the Public Affairs Office at KSC, the following concerns have been identified and are addresed by this project:

a. The process of obtaining information from large volumes of printed documentation is manual; it is cumbersome and not fast enough. An increase in media questions is expected.

b. There is a turnover in personnel and long training periods are required to learn how to properly answer media questions.

c. Answers to media questions must be saved and later searched to avoid different answers to the same question and to reduce duplication of effort.

d. An increase in the amount of documentation is expected.

e. Coordination with the Public Affairs operations within other NASA space centers is anticipated.

To summarize, the aim is to eliminate an employee turnover and training problem for a regularly performed, publicly visible task which is tedious, labor intensive, and can produce an inconsistent result.

Progress to date has been to research hardware and software requirements for a prototype intelligent interactive high speed data search system to automate the manner in which media questions are answered. The effort has led to a prototype design aimed at expository text (technical manuals, reports, regulations, research papers, etc.). The prototype takes advantage of IR techniques to find relevant text, and it uses general world knowledge to answer queries and summarize text. As diagrammed in Figure 1, it is characterized by the following three activities:

I. Automatic indexing based on syntactic, linguistic, and probabilistic techniques to retrieve a ranked list of relevant text for a given natural language query.

II. Converting retrieved relevant text to a conceptual format based on general or world knowledge.

III. Letting meaning be represented by "portions" of the original text and using portions of retrieved text to build a response to a query or generate a summary.

This design concerns natural language processing and involves the acquisition of general world knowledge. These are complex tasks still open to research and criticism. Consequently, some detail is necessary to make it clear as to what will be accomplished by this project.

Activity I is not new; there are commercial systems which already perform this activity. These systems are complex but rather straightforward to implement. They are based on frequency of occurrence of character strings and the more sophisticated systems know how words should be spelled and what part of speech a word can have (e.g., noun, verb, adjective, etc.). Referring to Figure 1, for a demonstration, the 1,000 page shuttle manual (stored electronically as a three megabyte text file) was used by considering each paragraph of the manual as a document. This resulted in a collection of 4,902 documents. A commercial hypertext system called SPIRIT was used to automatically index the collection and provide natural language access. For the natural language query "When does NASA select astronaut candidates?", a ranked list of forty-three relevant paragraphs was retrieved. Paragraph number 3386 was the most relevant, paragraph number 2132 was the second most relevant, etc. For the prototype system, a ranked list of paragraphs retrieved for a query is stored in RAM to facilitate further processing.

Activity II is based on a linguistic concept called thematic roles. Thematic roles help question answering by revealing how sentence phrases and clauses are related to the verbs and modifiers in a sentence. It is easy to explain thematic roles by using an example and avoiding some detail. Consider the following sentence:

Mary made coffee for John with a percolator.

There are four noun phrases in this sentence, each of

which fits into a particular thematic role as follows:

Noun Phrase	Thematic Role
Mary	AGENT
coffee	THEME
for John	BENEFICIARY
with a percolator	INSTRUMENT

Four corresponding questions can be answered now:

What was made? -> THEME -> coffee
Who made coffee? -> AGENT -> Mary
For whom was coffee made? -> BENEFICIARY -> John
With what was coffee made? -> INSTRUMENT -> a percolator

There are approximately thirty known thematic roles.

It is important to point out that this project is not concerned with figuring out, for example, what "a percolator" is; we only need to know that this string of characters occupies an INSTRUMENT position in the thematic form of a sentence. This is perhaps the most important point behind the success of this project.

In order to convert text to thematic form, the following general world knowledge is needed:

a. Certain words suggest specific thematic roles. For example, all verbs other than the "be-verbs" (is, are, was, etc.) become candidates for an ACTION. As another example, the word "noon" implies TIME.

b. Prepositions and certain conjunctions trigger thematic role possibilities. For example, the preposition "for" can trigger the BENEFICIARY or DURATION thematic roles. The conjunction "as" can possibly trigger the MANNER or TIME thematic roles.

c. Normally, verbs cause thematic grids of the form AGENT followed by ACTION as in

   John ran.

   or the form AGENT followed by ACTION followed by THEME as in

   John chased the dog.

   Each of these grids can then be followed by some of the other thematic roles as in

   John chased the dog into the house.

   Here, "into the house" is a DESTINATION thematic role. But there are other thematic grids. For example, there is a class of verbs (represented by the verbs "load" or "spray") which can have a grid which specifies an AGENT, ACTION, LOCATION, THEME order, as in

   John loaded the truck with furniture.

   Here, "the truck" is a LOCATION rather than a THEME.

The conversion process uses the knowledge outlined above for all words which may appear in sentences to be converted. For a somewhat robust system, this would involve approximately 100,000 words. Items B amd C are the most important with item C being the most time consuming to establish. The conversion process can operate even if knowledge for items A and C is not known for domain specific words; this makes the process domain independent.

Acquisition of the above knowledge will be done manually from memory, but also from reading English dictionary entries when we are not so sure. Creating this knowledge is a time consuming process. However, it is felt that the knowledge can be automatically produced directly from standard English dictionary text. We have only just begun to consider procedures for this task. In any case, the knowledge must be established and it can be obtained from an English dictionary. This is shown as an input to Activity II in Figure 1 and labeled as mass storage for common knowledge.

The output of Activity II is a collection of thematically parsed sentences stored in RAM to facilitate further processing. Referring to Figure 1, a thematic parse for the sentence "Astronaut candidates are selected as needed for training at JSC." is shown as an example. This sentence occurred in paragraph number 4011 of the shuttle manual, the third most relevant paragraph for the query "When does NASA select astronaut candidates?". For the forty-three retrieved paragraphs shown as input to Activity II, there are 160 sentences and potentially each one would have to be converted to thematic form.

Activity III is rather straightforward. An answer to the query is generated by a scan of parsed sentences for one having a TIME portion, where the ACTION specified is "select" and the THEME specified is "candidates"; and if an AGENT is present, it must have the value "NASA". The response is the TIME portion of such a sentence. In Figure 1, this is shown as the phrase "AS NEEDED" on the computer monitor, taken from the sentence mentioned above and diagrammed on the RAM device in Figure 1.

A crude version of the SPIRIT system's retrieval capabilities has been implemented in the C programming language for a PC or SUN workstation. Speed is not a problem for this activity; good performance can be achieved in a PC environment. A crude thematic parser is under development now to investigate speed performance in a PC environment for Activities II and III. The Prolog programming language is being used to enable rapid development of the parser. Parallel parsing techniques are also being considered because high speed appears to be an issue here.

The project's schedule includes a quickly implemented prototype incorporating all three activities by October 1990. This system will only have a 600 word vocabulary but should demonstrate relevant text retrieval and thematic parsing ability. Its purpose is to investigate the speed issue mentioned for Activity II. The current level of funding allows for manual development of a 14,000 word vocabulary and demonstration of a question/answer prototype in December 1991, and a summarization prototype in December 1992. These prototypes could handle, for example, the shuttle manual.

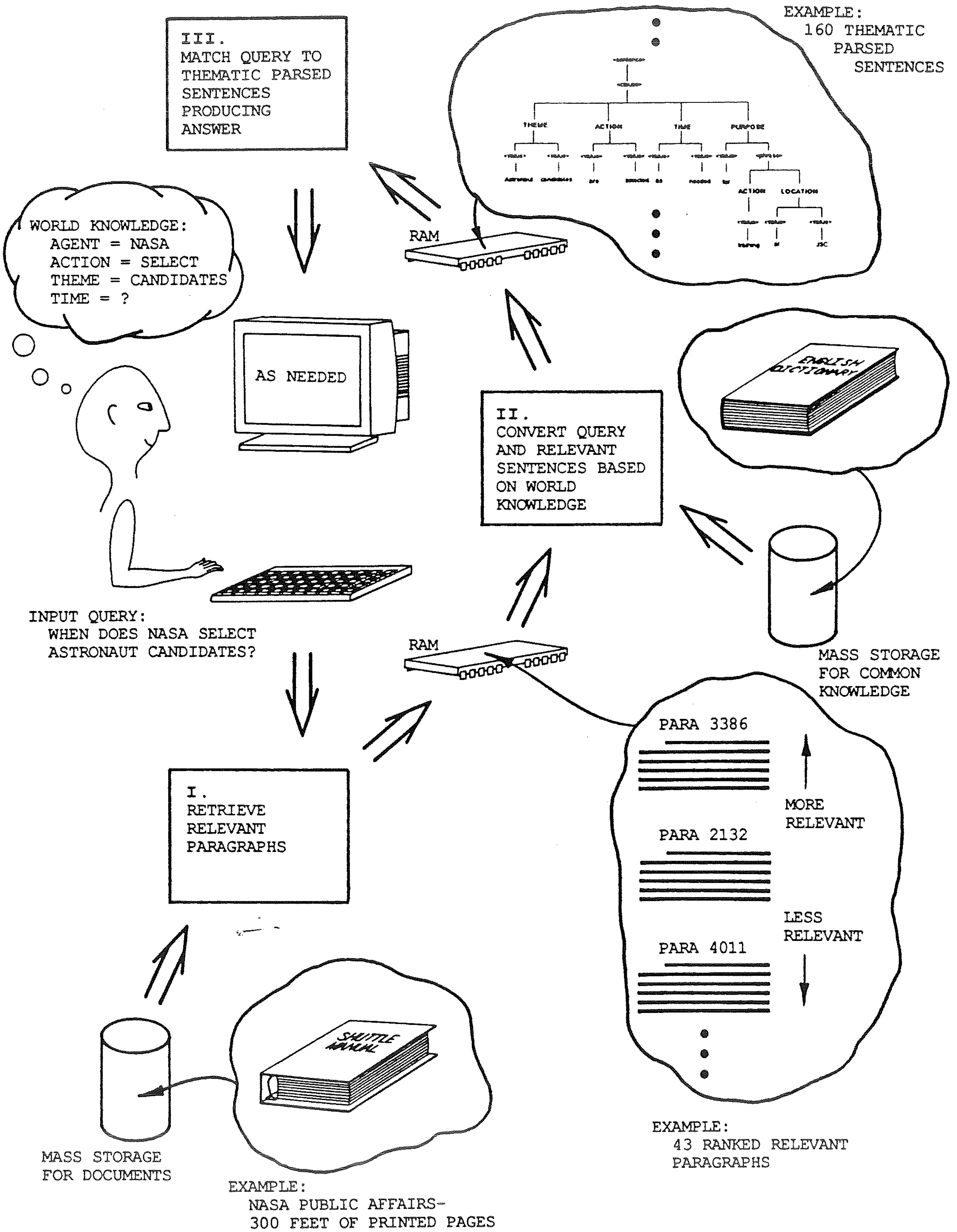


Figure 1: Three Activities for Prototype Question/Answer System.

Section 3.

Master's Degree Research Project Proposal submitted to Graduate Committee, Department of Computer Science, UCF

IMPLEMENTATION OF A THEMATIC ROLE AND ENTITY ATTRIBUTE LEXICON FOR INFORMATION RETRIEVAL FROM FULL-TEXT DATABASES

Research Project Proposal Submitted to:

Graduate Committee
Department of Computer Science
University of Central Florida

by

Edgar B. Wendlandt

Advisory Committee

Dr. J. R. Driscoll, Chairman
Dr. M. A. Bassiouni
Dr. S. D. Lang

November 13, 1990

I. Introduction

Full-text information retrieval (IR) systems differ from record based IR systems. Identifiers, also referred to as keywords, index terms, or descriptors, are attached upon entry of the text data . The identifiers may be attached manually or they may be automatically assigned. The automated process performs two functions of interest during the initial entry of the data. First, it throws out useless or empty words. These words are generally prepositions. Second, it calculates weighting factors. The weighting factors indicate how useful a document is to a query. Document size is variable and generally assigned by the user. For instance, each paragraph in a text may be considered a document. The calculation of the weighting factor *(w)* is a combination of term frequency *(tf)*, document frequency *(df)*, and inverse document frequency *(idf)*. The terms are defined as follows:

tf_{ij} = number of occurrences of term T_j in document D_i

df_j = number of documents in a collection which contain T_j

idf_j = $\log(N/df_j)$; where N = total number of documents.

w_{ij} = $tf_{ij} * idf_{ij}$

When the system is queried it computes a vector Q of weighting factors. The retrieval of a document is based on the value of a similarity coefficient. In current literature the degree of similarity is most often a factor of term frequencies, co-occurrence relationships, and proximal distances [DEB89,RAU89,ZER90]. Some common similarity factors are calculated as follows:

$$sim_1(Q,D_i) = \sum_{j=1}^{t} w_{qj} * d_{ij}$$

$$sim_2(Q,D_i) = \frac{\sum_{j=1}^{t} w_{qj} * d_{ij}}{\sqrt{\sum_{j=1}^{t} d_{ij}^2}}$$

$$sim_3(Q,D_i) = \frac{\sum_{j=1}^{t} w_{qj} * d_{ij}}{\sqrt{\sum_{j=1}^{t} d_{ij}^2 * \sum_{j=1}^{t} w_{qj}^2}}$$

The function sim_1 is a crude but simple coefficient. The functions sim_2 and sim_3 normalize the coefficients in case of different document sizes.

II. Objective

The objective of the project is to modify and potentially improve an existing IR system. The existing system is now being applied to Dr. Driscoll's NASA KSC funded research. The system currently uses inverted indexing and keyword weighting as described. In addition it has the ability to recognize synonyms of keywords.

As an example, given a query about the space shuttle program, the system currently aids the user in obtaining relevant documents, which are paragraphs from the space shuttle manual (a one-thousand page reference manual). Output from the current system is generally a long list of rank ordered paragraph numbers. The user must then scan these paragraphs for the factual information desired. Refer to Figure 1 where a sample query is shown at the top. The system returns a list of six empty words, seven keywords, and a listing of relevant paragraph classes. These classes are groups of paragraphs of similar weight where the number of paragraphs in a group is listed under NB DOCS. The keywords that triggered the retrieval of these paragraphs is listed to the right. The class of paragraphs are listed from most relevant to least relevant as determined by the present system. The only paragraph from class one is shown at the bottom of the figure. In this case, the answer to the query happens to be in that paragraph.

Refer to Figure 2, where the same system was used to answer 35 questions. (Appendix A provides a list of the questions.) The figure shows that it is not always the case that the first paragraph contains the answer to the question. In the figure, the right most column shows the number of paragraphs that were scanned in the order the system ranked them before the answer was found. The last line indicates that an average of three paragraphs had to be scanned to find the answer to the questions but notice that in a few cases, eleven paragraphs were scanned before the answer was found.

Shuttle Data Base: 5000 documents, each a paragraph in the shuttle manual

<1> : How long does the payload crew go through training before a launch?

EMPTY WORDS: how, does, the, through, before, a.
KEY WORDS: long, does, payload, crew, go, training, launch.

CLASSES	NB DOCS	KEY WORDS
1	1	Payload-crew, training, launch
2	2	payload-crew, training
3	1	long, payload, crew, training, launch
4	1	long, payload, crew, training
5	1	payload, crew, go, launch
6	1	does, go, launch
7	1	does, payload, training
8	1	does, crew, go
9	1	crew, go, training
10	1	long, payload, training
11	1	long, does, crew
12	1	does, payload, launch
13	3	payload, training, launch
14	2	crew, go, launch
15	1	does, crew, launch
16	2	crew, training, launch
17	3	payload, crew, go
18	17	payload, crew, training
19	2	long, crew, launch

Ignored Thematic Information:

1. Query
   How long ⇒ DURATION, ...
   through ⇒ DURATION
   before ⇒ TIME
2. Sentence Answer
   through ⇒ DURATION
   before ⇒ TIME

DOC 3386 BASE:doc 3386 NCP:0/CPI:1/NBI:1 +18 1K/1K
IDENTIFIER:doc 3386
TEXT.........:
Training requirements depend on the complexity of both the instruments and the integrated payload. The investigator helps determine training requirements for his instrument and participates in training the payload specialists and support personnel who may operate equipment, monitor data or assist in troubleshooting. Investigators who participate in such activities also require indoctrination and training in practices and equipment operation. Payload crew training currently begins 1.5 to two years before launch and continues through launch. Training on individual experiments and experiment instruments is normally scheduled before payload integration, which begins approximately one year before launch.

Figure 1. Sample Query

Results of finding answers using SPIRIT:
(ordered by question number)

	Answer found?	Search Time	Number of documents read
1)	found	1 minute 45 seconds	1
2)	found	1 minute 45 seconds	1
3)	found	1 minute 45 seconds	2
4)	found	2 minutes 0 seconds	3
5)	found	1 minute 30 seconds	2
6)	found	2 minutes 0 seconds	3
7)	found	1 minute 30 seconds	1
8)	found	3 minute 0 seconds	3
9)	found	1 minute 15 seconds	1
10)	found	1 minute 15 seconds	1
11)	found	4 minutes 30 seconds	11
12)	found	1 minute 0 seconds	1
13)	found	1 minute 0 seconds	1
14)	found	2 minutes 45 seconds	6
15)	found	2 minutes 30 seconds	6
16)	found	1 minute 0 seconds	1
17)	found	1 minute 30 seconds	1
18)	found	1 minute 0 seconds	1
19)	found	2 minutes 0 seconds	2
20)	found	2 minutes 15 seconds	5
21)	found	1 minute 0 seconds	1
22)	found	2 minutes 0 seconds	1
23)	found	3 minutes 0 seconds	5
24)	found	1 minute 30 seconds	1
25)	found	3 minutes 30 seconds	11
26)	found	3 minutes 45 seconds	8
27)	found	1 minute 15 seconds	1
28)	found	2 minutes 0 seconds	2
29)	found	3 minutes 0 seconds	4
30)	found	2 minutes 45 seconds	6
31)	found	3 minutes 30 seconds	5
32)	found	1 minute 15 seconds	1
33)	found	1 minute 45 seconds	1
34)	found	1 minute 15 seconds	1
35)	not found	1 minute 0 seconds	0*

*Could determine in one minute that answer was not in the book. Notice that the word "color" was not found as a keyword.

Average search time = 2 minutes 2 seconds

Average number of documents read = 3

Figure 2. Query Results

It has been suggested that thematic roles and entity attributes may be used to better extract the desired document [SAL89,ZER90]. Thematic roles are a linguistic concept. They could help in question answering by revealing how sentence phrases and clauses are related to the verbs and modifiers in a sentence. There are approximately thirty known thematic roles. Entity attributes are generally considered descriptors of something. They are important in answering questions about an object's attributes. For example, color and weight are considered attribute categories.

In this respect, the current system throws out some useful information when it blindly throws out the empty words. Again, refer to Figure 1 where the box out to the right shows ignored thematic information. The box shows that the thematic information of duration and time were ignored when the words "how," "through," and "before" were thrown out as empty words. Also, it shows that the sentence containing the answer has thematic information of duration and time.

A major obstacle in the use of thematic roles and entity attributes is the non-availability of an appropriate lexicon [ZER90]. Therefore, a thematic role and entity attribute lexicon must be built. The goal is to implement a system that uses the lexicon to better place the most relevant paragraphs relating to a query at the head of the list.

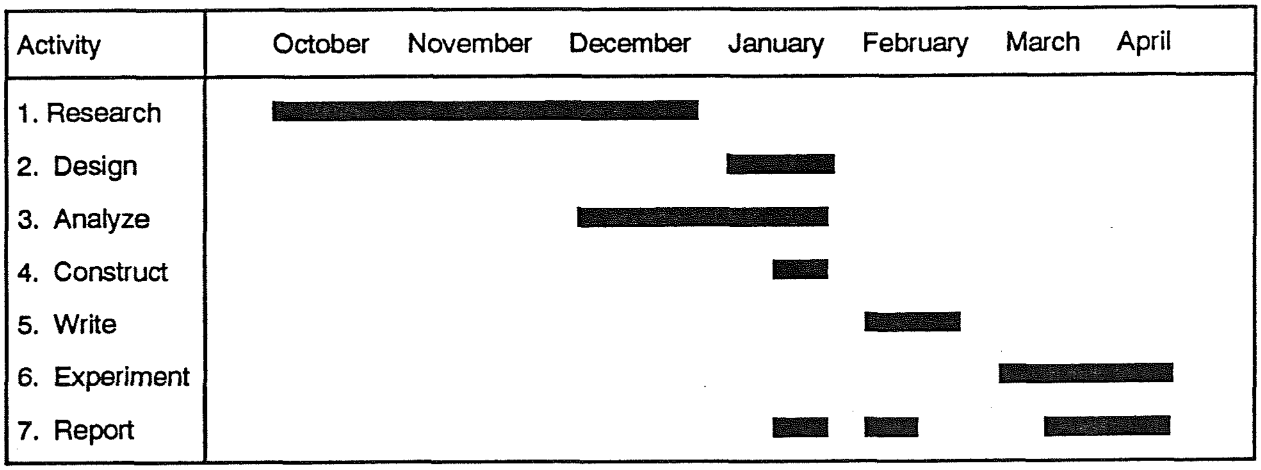
III. Activities and Significance

The first portion of the project requires designing and manually building a limited lexicon for a text file containing 26 pages of the NASA Shuttle Manual. After the lexicon has been designed and constructed, the existing IR system will be modified to incorporate the use of thematic roles and attributes. Weighting factors for thematic role and entity attribute occurrence will be computed in a fashion similar to the weighting factors mentioned in Section I. This new weighting factor will be combined with the existing weighting factor in similarity coefficient calculations. Details of the combination will be determined as an activity for this project.

In the second phase of the project, a series of experiments will be run. The output of the modified system will be compared to that of the original system. Depending on the results of the initial experiments the newly introduced weighting factors may be proportioned to place more or less significance on the thematic roles and attribute characteristics.

Finally, the results of the experimentation phase will be evaluated. The evaluation will provide some feedback about the lexicon's effectiveness and content. This feedback may aid in developing algorithms for general lexicon construction. Also, it should indicate whether or not the concept of thematic roles is significant in the area of fact retrieval from full-text databases.

IV. Chronology



1. Research of current techniques used in IR and submit proposal.
2. Design the lexicon and algorithms for incorporating its use (4 weeks).
3. Analyze the source code of the current IR system (10 weeks).
4. Construct a limited domain lexicon (2 weeks).
5. Write and implement the source code modifications (4 weeks).
6. Experiment, result analysis (8 weeks).
7. Report Preparation.

V. Detailed Chronology

Activity 1: The research has been ongoing throughout the fall semester.

Activity 2: The design of the lexicon will require specifying how the lexicon will be stored on the computer. The algorithm design will specify how the lexical information will be incorporated.

Activity 3: The IR system to be modified is a C program written by students under the guidance of Dr. Driscoll. I need to determine how and where functions may be added to modify the weighting to reflect the addition of lexical information.

Activity 4: The building of the lexicon will require that I manually scan the test document noting words with thematic and attribute information. Also, a set of words "typically" used in queries will need to be determined. These words and their associated thematic and attribute information will also need to be included in the lexicon.

Activity 5: The C source code will need to be written for the functions designed in Activity 2. The current IR system source will also be modified to incorporate these functions as determined in Activity 3.

Activity 6: The experiment will require running the original system and the new system on some test queries. The results will be analyzed to determine if the thematic and attribute information was incorporated as best as possible and if not modifications may be made requiring additional runs.

Activity 7: The report preparation will be ongoing. It will include a summation of how the lexicon was constructed, a summation of exactly how the thematic and attribute information was incorporated, an analysis of the results, and conclusions.

VI. List of Deliverables

1. Design document and the algorithms for incorporating lexicon.
2. Limited Domain Lexicon.
3. Design document and the source listing for the IR system modification routines.
4. Experiment results.
5. Project report.

VII. BIBLIOGRAPHY

[DEB89] Debili, Fathi, "About Reformulation in Full-Text IRS," Information Processing and Management (Vol 25), 1989, pp. 647-657.

[KEM88] Kemp, Alasdair D., Computer-Based Knowledge Retrieval, The Association for Information Management, 1988, pp. 4-1 - 4-29.

[RAU88] Rau, Lisa F., "Conceptual Information Extraction and Retrieval From Natural Language Input," RAIO (Vol 1), 1988, pp. 424-437.

[RAU87] Rau, Lisa F., "Knowledge Organization and Access in a Conceptual Information System," Information Processing & Management (Vol 23 No 4), 1987, pp. 269-283.

[RAU 89] Rau, Lisa F., Paul S. Jacobs, Uri Zernik, "Information Extraction and Text Summarization Using Linguistic Knowledge Acquisition," Information Processing & Management (Vol 25 No 4), 1989, pp. 419-428.

[SAL89] Salton, Gerard, Automatic Text Processing, Addison Wesley, 1989, pp. 275-375.

[ZER90] Zernik, Uri, and Paul Jacobs, "Tagging for Learning: Collecting Thematic Relations from Corpus," COLING-90 (Vol 1), 1990, pp. 34-39.

Appendix A

Shuttle Questions:

1) What is the maximum cargo weight the shuttle can carry?
2) How far can the shuttle transport cargo from the earth's surface?
3) What has happened to the Enterprise?
4) How many years of education are required for astronaut candidates?
5) What is the total weight of the shuttle?
6) How thick is the window of the shuttle?
7) How many gallons of liquid hydrogen fuel can the storage tank hold?
8) What type of liquid fuel is used on the shuttle?
9) What is the descent rate of the shuttle during landing?
10) How long is the mechanical arm used for payload deployment?
11) What are the dimensions of the cargo area in the shuttle?
12) How is waste disposed of?
13) Have there been astronauts picked from minority groups?
14) What is the total number of times that the shuttle has been launched?
15) What type of food do astronauts eat during a shuttle mission?
16) What is the orbiter's velocity while in orbit?
17) What is the maximum acceleration of the shuttle during launch?
18) What is the maximum touchdown glide speed of the shuttle?
19) How many pounds of thrust do the SRB booster rockets generate during liftoff?
20) What is the maximum fluid fuel flow rate during launch?
21) How fast does the crawler or transporter travel?
22) At what altitude and speed must the pilot throttle back during ascent?
23) How many general purpose computers are on board the shuttle and what functions do they serve?
24) What is the new design of general purpose computers like on board the shuttle?
25) What is the total number of tiles that cover the orbiter for thermal protection during reentry?
26) When did the first space shuttle launch occur?
27) How long is the runway at a shuttle landing facility?
28) What type of glass is used for the windows to withstand the pressure of flight?
29) What is the total amount of RAM available in the shuttle's general purpose flight computer?
30) What material are the heat shield tiles composed of?
31) What type of computer guidance and navigation system does the shuttle use during reentry and landing?
32) What is the maximum power available to the payload area?
33) Are there emergency escape procedures to jettison the crew members out of the shuttle?
34) Where do the crew members sleep on the shuttle?
35) What is the color of the external tank? (no answer)

Section 4.

Proposal funded by State of Florida,

High Technology and Industry Council (FHTIC)

(Note to SUS DSR, this is top page of submitted proposal)

A Proposal
submitted to

State of Florida
HIGH TECHNOLOGY AND INDUSTRY COUNCIL
APPLIED RESEARCH GRANTS PROGRAM

by

UNIVERSITY OF CENTRAL FLORIDA
Division of Sponsored Research
P. O. Box 25000 ADM 243
Orlando, Florida 32816-0001
(407) 275-2671

Title: Intelligent Text Processing

Planning: X Operational: ______ No. of years funded to date: 0

Technology Area: Software and Computer Science

Principal Investigator(s):

Name: James R. Driscoll

Signature: J. Driscoll

Telephone: (407) 275-2341

Soc. Sec. No. [redacted]

Dept. or Unit Affiliation: Dept. of Computer Science

Institution (for consortium members):

Proposal: New X Renewal ______ 1990-91 Amount Requested: $ 19,986.00

Executive Summary (100 words): This proposal builds upon ongoing externally funded reserach in regard to development of a prototype, PC-based, intelligent high speed text data search system. This proposal has two objectives. The first objective is to develop a fully-functional, PC-based text retrieval system incorporating the very latest proven linguistic techniques. We expect this first objective to yield a commercial product. The second objective is to plan for commercialization of more intelligent text processing systems for such tasks as answering questions, writing summaries, correcting grammar, or translating text from one language to another.

Endorsements:

	Department Head (or Dept. Center Director)	College Official (or Multi-Dept. Center Director)	University Official
Name	Dr. Ron Dutton	Dr. Bruce A. Whisler	Dr. Joan R. Burr
Signature	Ronald Dutton	Bruce A. Whisler	Joan R. Burr
Title	Assoc. Chairman	Assistant Dean for Budget	Director, Division of Sponsored Research
Telephone			
Date			9/10/90

TABLE OF CONTENTS

INTRODUCTION

Everyone is familiar with the commercial success of word processors and their presence on virtually every desk in an office environment. These systems are invaluable for creating documents, and often appear intelligent when they perform tasks such as correcting spelling. Now, an era of commercially available text processing programs is approaching. For example, we can expect to see, in the not too distant future, intelligent text processors which can correct grammar, search existing text and answer questions, scan existing text and write summaries, and automatically translate text to other languages.

It is reasonable to expect these products to be just as prevalent as word processors are today. Several factors are contributing to this expectation. One is the advance in storage and computing power of "desktop" computers. Another is the ability to store and process large machine readable dictionaries and thesauri. And a third is the ability to store and process massive corpora of "everyday" text.

These advances and abilities have triggered new linguistic research. Automatically parsing sentences, automatically obtaining semantic information from dictionaries, and machine translation were the most prevailing research topics presented at the 13th International Conference on Computational Linguistics held August 16-25, 1990. In addition, intelligent text handling has become a rapidly developing research field in its own right. Refer to the Call for Papers and Call for Demonstrations at the RIAO 91 Conference on Intelligent Text and Image Handling in ADDITIONAL SUPPORTING DOCUMENTATION.

Still, there are some significant research hurdles for intelligent text processing. The TECHNICAL DISCUSSION which follows describes a project at the University of Central Florida (UCF) for developing a prototype system for natural language access to large amounts of text. The research has been supported by DoD Army Research Institute/PM-Trade Contract

N61339-88G-0002 Order 0004,[1] and by NASA Kennedy Space Center (KSC) Grant NAG 10-0058 Project 2A.[2] This ongoing project at UCF attempts to solve a major text processing research hurdle and several facets of this funded research could lead to commercialization.

TECHNICAL DISCUSSION

1. Detailed Statement of the Problem

This project concerns natural language access to large amounts of text. The objective is to automate the access of textual information so that users are not overwhelmed by the amount of information searched. Currently, the project is motivated by the desire to provide convenient access to information contained in the numerous and large public information documents maintained by Public Affairs at KSC.

The documents maintained by Public Affairs at KSC consist of press releases and other printed information created at KSC and other NASA offices using various word processors. There are also documents from outside contractors, such as Rockwell, which produces the "NASA National Space Transportation System Reference" more often called the "shuttle manual." During a launch at KSC, about a dozen NASA employees access these printed documents to answer media questions. The amount of text is visually overwhelming. The shuttle manual alone is just over 1,000 pages (it is three inches thick). The planned document storage for NASA KSC Public Affairs is about 100 million words of text or around 300,000 pages (approximately seventy-five feet of stacked pages) with about 5,000 pages added or replaced annually. Electronically, the storage is manageable; approximating 900 megabytes of disk space is expected.

---

1 This research demonstrated the usefulness of thematic roles (surface knowledge) in automated diagnostic systems. Thematic roles were used to match a user's description of a problem with symptoms, to obtain procedures to correct the problem.

2 This research provided a survey of hardware and software technology appropriate to automating the manner in which NASA KSC Public Affairs personnel obtained answers to media questions. It resulted in a demonstration of a commercial state-of-the-art text retrieval system to NASA KSC.

In regard to the operation of the Public Affairs Office at KSC, the following concerns have been identified and are addressed by this project:

a. The process of obtaining information from large volumes of printed documentation is manual; it is cumbersome and not fast enough. An increase in media questions is expected.

b. There is a turnover in personnel and long training periods are required to learn how to properly answer media questions.

c. Answers to media questions must be saved and later searched to avoid different answers to the same question and to reduce duplication of effort.

d. An increase in the amount of documentation is expected.

e. Coordination with the Public Affairs operations within other NASA space centers is anticipated.

To summarize, the aim is to eliminate an employee turnover and training problem for a regularly performed, publicly visible task which is tedious, labor intensive, and can produce an inconsistent result.

2. Progress

Progress to date has been to research hardware and software requirements for a prototype intelligent interactive high speed data search system to automate the manner in which media questions are answered. The effort has led to a prototype design aimed at expository text (technical manuals, reports, regulations, research papers, etc.). The prototype takes advantage of IR techniques to find relevant text, and it uses general world knowledge to answer queries and summarize text. As diagrammed in Figure 1, it is characterized by the following three activities:

I. Automatic indexing based on syntactic, linguistic, and probabilistic techniques to retrieve a ranked list of relevant text for a given natural language query.

II. Converting retrieved relevant text to a conceptual format based on general or world knowledge.

III. Letting meaning be represented by "portions" of the original text and using portions of retrieved text to build a response to a query or generate a summary.

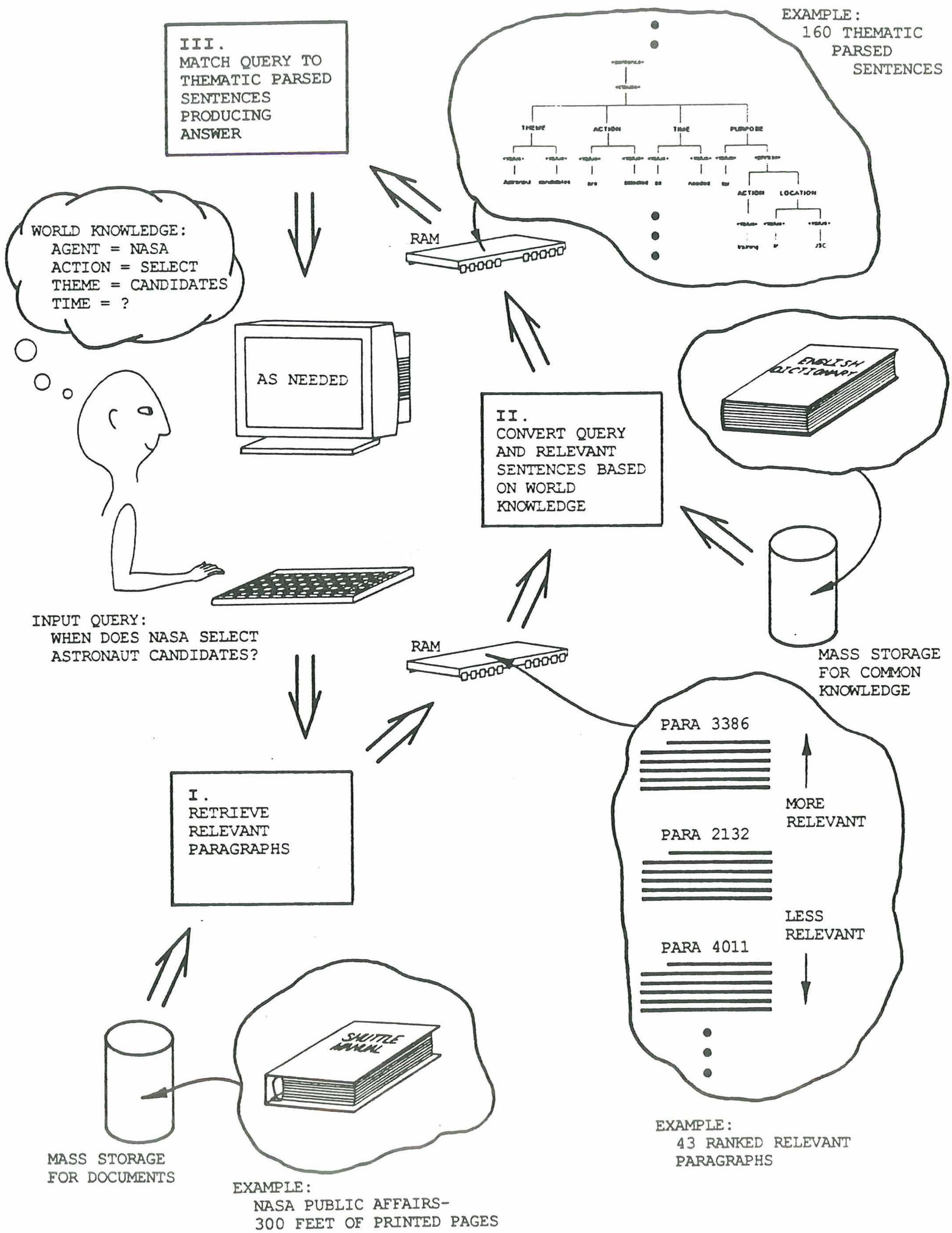


Figure 1: Three Activities for Prototype Question/Answer System.

This design concerns natural language processing and involves the acquisition of general world knowledge. These are complex tasks still open to research and criticism. Consequently, some detail is necessary to make it clear as to what will be accomplished by this project.

3. Method of Attack

Activity I is not new; there are commercial systems which already perform this activity. These systems are complex but rather straightforward to implement. They are based on frequency of occurrence of character strings and the more sophisticated systems know how words should be spelled and what part of speech a word can have (e.g., noun, verb, adjective, etc.). Referring to Figure 1, for a demonstration, the 1,000 page shuttle manual (stored electronically as a three megabyte text file) was used by considering each paragraph of the manual as a document. This resulted in a collection of 4,902 documents. A commercial hypertext system called SPIRIT[3] was used to automatically index the collection and provide natural language access. For the natural language query "When does NASA select astronaut candidates?" a ranked list of forty-three relevant paragraphs was retrieved. Paragraph number 3386 was the most relevant, paragraph number 2132 was the second most relevant, etc. For the prototype system, a ranked list of paragraphs retrieved for a query is stored in RAM to facilitate further processing.

Activity II is based on a linguistic concept called thematic roles. Thematic roles help question answering by revealing how sentence phrases and clauses are related to the verbs and modifiers in a sentence. It is easy to explain thematic roles by using an example and avoiding some detail. Consider the following sentence:

Mary made coffee for John with a percolator.

3 Refer to the DATABASE journal article about SPIRIT in ADDITIONAL SUPPORTING DOCUMENTS.

There are four noun phrases in this sentence, each of which fits into a particular thematic role as follows:

Noun Phrases	Thematic Roles
Mary	**AGENT**
coffee	**OBJECT**
for John	**BENEFICIARY**
with a percolator	**INSTRUMENT**

Four corresponding questions can be answered now:

What was made ? → **OBJECT** → coffee
Who made coffee? → **AGENT** → Mary
For whom was coffee made ? → **BENEFICIARY** → John
With what was coffee made ? → **INSTRUMENT** → a percolator

There are approximately thirty known thematic roles.

It is important to point out that this project is not concerned with figuring out, for example, what "a percolator" is; we only need to know that this string of characters occupies an **INSTRUMENT** position in the thematic form of a sentence. This is perhaps the most important point behind the success of this project.

In order to convert text to thematic form, the following general world knowledge is needed:

a. Certain words suggest specific thematic roles. For example, all verbs other than the "be-verbs" (is, are, was, etc.) become candidates for an **ACTION**. As another example, the word "noon" implies **TIME**.

b. Prepositions and certain conjunctions trigger thematic role possibilities. For example, the preposition "for" can trigger the **BENEFICIARY** or **DURATION** thematic roles. The conjunction "as" can possibly trigger the **MANNER** or **TIME** thematic roles.

c. Normally, verbs cause thematic grids of the form **AGENT** followed by **ACTION** as in

   John ran.

   or the form **AGENT** followed by **ACTION** followed by **THEME** as in

   John chased the dog.

   Each of these grids can then be followed by some of the other thematic roles as in

   John chased the dog into the house.

   Here, "into the house" is a **DESTINATION** thematic role. But there are other thematic grids. For example, there is a class of verbs (represented by the verbs "load" or "spray") which can have a grid which specifies an **AGENT, ACTION, LOCATION, THEME** order, as in

John loaded the truck with furniture.

Here "the truck" is a **LOCATION** rather than a **THEME**.

The conversion process uses the information outlined above for all words which may appear in sentences to be converted. For a somewhat robust system, this would involve approximately 100,000 words. Items B and C are the most important with item C being the most time-consuming to establish. The conversion process can operate even if knowledge for items A and C is not known for domain specific words; this makes the process domain independent.

Acquisition of the above knowledge will be done manually from memory, but also from reading English dictionary entries when we are not so sure. Creating this knowledge is a time consuming process. However, it is felt that the knowledge can be automatically produced directly from standard English dictionary text. We have only just begun to consider procedures for this task. In any case, the knowledge must be established and it can be obtained from an English dictionary. This is shown as an input to Activity II in Figure 1 and labeled as mass storage for common knowledge.

The output of Activity II is a collection of thematically parsed sentences stored in RAM to facilitate further processing. Referring to Figure 1, a thematic parse for the sentence "Astronaut candidates are selected as needed for training at JSC" is shown as an example. This sentence occurred in paragraph number 4011 of the shuttle manual, the third most relevant paragraph for the query "When does NASA select astronaut candidates?" For the forty-three retrieved paragraphs shown as input to Activity II, there are 160 sentences and potentially each one would have to be converted to thematic form.

Activity III is rather straightforward. An answer to the query is generated by a scan of parsed sentences for one having a **TIME** portion, where the **ACTION** specified is "select" and the **THEME** specified is "candidates"; and if an **AGENT** is present, it must have the value "NASA." The response is the **TIME** portion of such a sentence. In Figure 1, this is shown as the phrase "**AS NEEDED**" on the computer monitor, taken from the sentence mentioned above and diagrammed on the RAM device in Figure 1.

A crude version of the SPIRIT system's retrieval capabilities has been implemented in the C programming language for a PC or SUN workstation. We have found that speed is not a problem for this activity; good performance can be achieved in a PC environment. A crude thematic parser is under development now to investigate speed performance in a PC environment for Activities II and III. The Prolog programming language is being used to enable rapid development of the parser. Parallel parsing techniques are also being considered because high speed appears to be an issue here.

The project's three year schedule includes a quickly implemented prototype incorporating all three activities by October 1990. This system will only have a 1000 word vocabulary but should demonstrate relevant text retrieval and thematic parsing ability. Its purpose is to investigate the speed issue mentioned for Activity II. The current level of funding allows for manual development of a 14,000 word vocabulary and demonstration of a question/answer prototype in December 1991, and a summarization prototype in December 1992. These prototypes could handle, for example, the shuttle manual.

4. Relevance to FHTIC Applied Research Grants Objectives

The system called SPIRIT which we used for Activity I in the section Method of Attack is a commercial product available on a PC for \$10,000 and on a mainframe computer for \$80,000. The system was loaned to the P.I. for demonstration to NASA KSC. Refer to ADDITIONAL SUPPORTING DOCUMENTATION for more information about SPIRIT.

SPIRIT was developed in France. The English version of SPIRIT is not as powerful as the French version. For example, there is no synonym capability in the English version. In addition, the user's manual for the system is in French! It is important to note that we have not been able to find a commercial system in the United States as convenient and powerful as SPIRIT.

The crude version of the SPIRIT system's text retrieval capabilities which we have implemented for the NASA KSC prototype system incorporating all three activities works okay;

but we know now that it should have synonym capability. We feel that addition of the synonym capability and a few other standard information retrieval concepts will make the NASA KSC prototype text retrieval system better than the French commercial product SPIRIT when it comes to English text databases. We also feel that some new thematic research related to that discussed for Activity II and Activity III of the NASA KSC prototype can also be incorporated within the retrieval system to make it truly better than the SPIRIT commercial product.

4.1 First Objective

Our first objective, then, is to carry out this relatively straightforward enhancement and start creating a robust text retrieval system with the intent to make it a commercial product. The planned enhancements, beyond features required for our NASA KSC prototype, can be carried out by one graduate student working half-time for one year and with minimum P.I. supervision. The synonym capability will be incorporated within the NASA KSC prototype before the proposed FHTIC funding period. The graduate student's main activities will be to add thematic techniques, fine tune the system, and create a user's manual. All this is indicated in the Schedule section and reflected in the BUDGET.

We plan to demonstrate the text retrieval system using NASA KSC text data by November, 1991; so several trips to KSC are planned to collect text files. We also plan to demonstrate the system at the Hypertext 91 Conference during November, 1991 using the text from the proceedings of the Hypertext 87, 89, and 91 Conferences.

4.2 Second Objective

The P.I. recently attended the 13th International Conference on Computational Linguistics. This year, the conference technical topics favored automated linguistic analysis using machine readable dictionaries, thesauri, and large common text corpora. It was learned at the conference that much of the text processing research required for completion of Activity II and Activity III

of the NASA KSC funded project is applicable to other endeavors such as machine translation and tools for linguistic analysis. Note that the NASA KSC funded project dictates the development of prototype systems for answering questions and summarizing text.

Consequently, our second objective is to plan for commercialization of the above mentioned text retrieval system and more intelligent text processing applications such as the NASA KSC funded question answer system (prototype available December, 1991), the NASA KSC funded text summarization system (prototype available December, 1992), a grammar correction system, a machine translation system, a toolbox of linguistic analysis programs, etc. The Schedule section and BUDGET reflect 10% release time for the P.I. to develop a comprehensive plan, and outline the expected return. Note that the NASA KSC approved funding as shown on the Schedule for years 1990, 1991, and 1992 represents a significant amount of matching funds for prototype development of these intelligent text processing systems.

5. Schedule

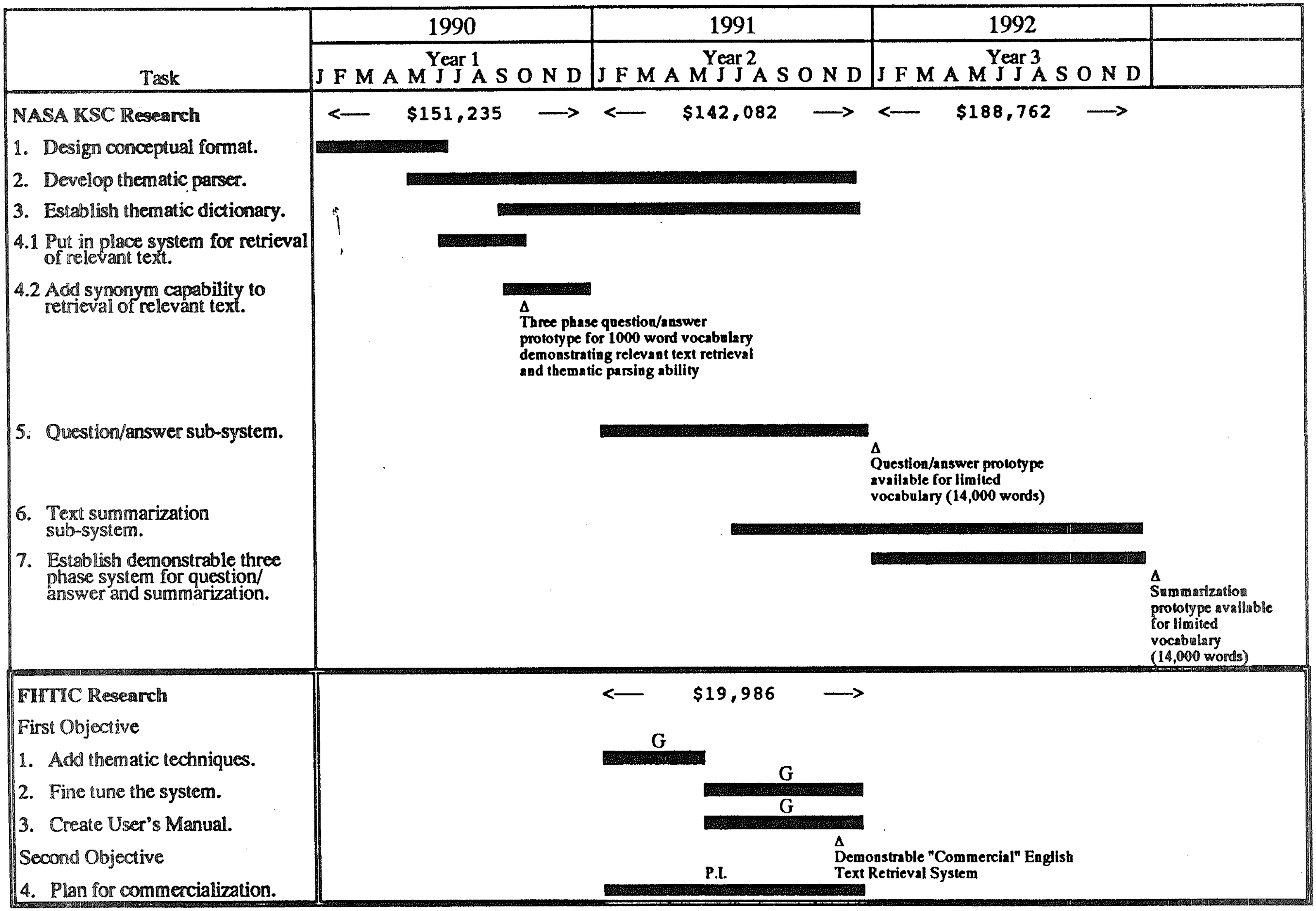
Refer to the chart on the next page for the schedule of activities for this project and its relationship to NASA KSC funded activities.

6. Project Summary

6.1 Technical Approach

The Army Research Institute PM-Trade provided $27,321 for Fall of 1988 to demonstrate the usefulness of thematic roles (surface knowledge) in regard to text processing. NASA KSC provided $33,209 during Summer and Fall of 1989 for the purpose of assessing the situation regarding high speed intelligent text retrieval. A survey of existing technology and commercial systems was performed. The results revealed that Europe is somewhat ahead of the United States in regard to commercialization of text retrieval systems and the application of linguistic research.

SCHEDULE



G = Graduate Student P.I. = Principal Investigator

The currently funded NASA KSC project described in the section Method of Attack makes use of established information retrieval techniques and new linguistic techniques. The linguistic techniques (thematic roles) appear to be consistent with new linguistic strategies and trends recently reported in linguistic journals and especially as reported at the 13th International Conference on Linguistics in Helsinki, Finland, August 15-25, 1990.

Because of all this, we are confident that the technical approach to the NASA KSC project is sound. Please note that NASA KSC has committed close to $500,000 for the demonstration of those existing and new text processing concepts in order to develop a PC-based prototype question answering system and a prototype text summarization system. These systems have application outside the space domain and, indeed, that is one reason why NASA has funded the project. In summary, the P.I. and several graduate students have already put two years of funded research effort into understanding the problem of intelligent text processing and planning for future prototype intelligent text processing systems.

6.2 Qualifications

In the text processing and database technology areas, the P.I. has demonstrated numerous international publications in the past four years, a significant amount of external funding ($627,000), and the ability to develop useful software systems. Furthermore, the P.I. has demonstrated international recognition as being knowledgeable in the area of intelligent text processing; note that he is on the program committee for the RIAO '91 International Conference on Intelligent Text and Image Handling (refer to ADDITIONAL SUPPORTING DOCUMENTATION).

The Department of Computer Science offers three degrees - B.S., M.S., and Ph.D. The department has seven hundred undergraduate majors and 150 graduate students. For 1988-1989, the department obtained external funds in the amount of $2,759,107 which included (1) an award of $1,903.131 which is evenly divided between a Computer Science faculty member, an Electrical Engineering faculty member, and a staff member from the Institute for Simulation and Training,

and (2) awards totalling $966,000 by two Computer Science faculty members working with the College of Engineering. The Computer Science department has access to numerous computing facilities; but note that NASA KSC funds provides all the computer hardware and software needed for the research proposed here.

6.3 Commercialization

As mentioned, most of the proposed effort in this first year of funding is aimed at development of a PC-based text retrieval system similar to a successful French commercial mainframe product which is not easily adaptable for use in the United States. Planning and assessment is proposed for more intelligent text processing software. After this first year of funding, the PC-based text retrieval system will exist and be demonstratable; only its promotion and marketing as a commercial product will remain. This may occur via an industrial partner or through UCF small business or marketing channels. A similar scenario is expected for the more intelligent text processing software such as the question answering system or text summarization system for which NASA KSC funds are being used to develop prototypes.

6.4 External Support for the Project

External support for this project already exists and is detailed in section 6.1 above. Via the NASA KSC funded research, various NASA subcontractors have heard about our research efforts and have expressed an interest in using the prototypes we are developing.

6.5 Florida University Infrastructure

At this point in time, no cooperative activities with other Florida universities exist. But an attempt will be made during the planning phase to identify possible cooperative activities. In regard to development of students for Florida industry, one particular NASA KSC subcontractor has been attending our project presentations and has requested the P.I. to identify students seeking employment who have experience related to the project.

FLORIDA HIGH TECH COUNCIL
APPLIED RESEARCH GRANTS PROGRAM
BUDGET FORM

Principal Investigator/Program Director: Dr. James R. Driscoll

DETAILED BUDGET FOR 12 MONTH PERIOD				From		Through
				Jan. 1, 1991		Dec. 31, 1991
Personnel		**Time/Effort**		**Dollar Amount Requested**		
Name	Postition/Title	%	Hours per week	Salary	Fringe Benefits	Totals
Dr. James R. Driscoll	P.I.	10	4	6,236.00	1,692.00	7,928.00
TBA	Graduate Student #1	50	20	11,440.00	18.00	11,458.00
			Subtotals	17,676.00	1,710.00	19,386.00

Equipment	
Supplies Computer disks, paper and copies	100.00
Other Expenses	
Travel Local travel to KSC (25 trips @ 100 mi. @ $.20/mi)	500.00
Consultant Costs	
Other Miscellaneous Costs	
Total Costs	$19,986.00

PRINCIPAL INVESTIGATOR/PROGRAM DIRECTOR: Dr. James R. Driscoll

Budget for Entire Project Period

In order to gain a funding history for this grant, indicate actual budget for each year funded and/or proposed. Also check if "Funded" or "Proposed" for each year (ie, new proposals will be "Proposed" for all years).

Budget CategoryTotals	1st Budget Period	Additional Years of Support			
		2nd	3rd	4th	5th
Personnel (Salary and fringe benefits)	$ 19,386.00				
Equipment					
Supplies	$ 100.00				
Other Expenses					
Travel	$ 500.00				
Consultant Cost					
Other Miscellaneous Costs					
Total Costs	$ 19,986.00				
Budget Status (✓)	___ Funded ___ Proposed	___ Funded ___ Proposed	___ Funded ___ Proposed	___ Funded ___ Proposed	___ Funded ___ Proposed
Total Costs For Entire Proposed Project Period					$ 19,986.00

Budget Justification:

1. The salary for Dr. Driscoll reflects (1) extra supervisory activity and (2) extra fact finding activity beyond that required for the existing NASA KSC project.
2. The salary for graduate student #1 is for implementation effort beyond that required under the existing NASA KSC project.
3. All computer hardware and software is available via the existing NASA KSC funded project.
4. Supply expenses are for those beyond the existing NASA KSC project.
5. Travel to conferences and other expenses is already provided by the existing NASA KSC project.
6. Approximately twenty-five trips to KSC beyond the existing NASA KSC project requirements are expected.

PERSONNEL

Name:	James R. Driscoll
Title:	Associate Professor of Computer Science
Birth Date:	January 18, 1948
Address:	Department of Computer Science University of Central Florida P. O. Box 25000 Orlando, Florida 32816
Home:	8380 Roanne Drive Orlando, Florida 32817
Telephone:	(407) 275-2341 (office) (407) 677-1108 (home)

EDUCATION

University of Kansas	B.S.	1971	Electrical Engineering (with Highest Distinction)
University of Kansas	M.S.	1974	Computer Science (with honors)
University of Kansas	Ph.D.	1977	Computer Science

M.S. Thesis: "Design of a Run-Time System for ALGOL 60"
D. Dissertation: "Towards the Design and Implementation of a Unified Data Based Management System"

PROFESSIONAL EXPERIENCE

Associate Professor, Department of Computer Science, University of Central Florida, 1981 to present.

Assistant Professor, Department of Computer Science, University of Central Florida, 1976-1981.

Instructor, Department of Computer Science, University of Kansas, 1974-76.

Graduate Teaching Assistant, Department of Computer Science, University of Kansas, 1972-74.

PROFESSIONAL DEVELOPMENT

Data Base Advisor, Institute for Simulation and Training, University of Central Florida, May 1986 to May 1988. Artificially Intelligent Keywording System.

Consultant, Insystec Corporation, Orlando, Florida, 1981-82. Data Base Management Systems.

Consultant, Harris Corporation Controls Group, Melbourne, Florida, 1977-78. Data Base Control of Power Systems.

RESEARCH IN PROGRESS

Development of a domain independent natural language interface to document databases, a domain independent diagnostic system, and domain independent systems for document abstracting, summarization, merging, language translation, and grammar correction.

RELEVANT PUBLICATIONS

L. C. Malone, J. R. Driscoll and J. W. Pepe, "Modeling the Performance of an Automated Keywording System" to appear in the journal Information Processing and Management.

R. Driscoll, D. A. Rajala, W. H. Shaffer and D. W. Thomas, "The Operation and Performance of an Artificially Intelligent eywording System," to appear in the journal Information Processing and Management.

/ C. Malone, J. W. Pepe and J. R. Driscoll, "Evaluation of an Automated Keywording System" Microcomputers for Information gement, June 1990.

K. Ray and J. R. Driscoll, "New Directions for Microcomputer-based Hypertext Systems" DATABASE Magazine, June, 1990.

L. Bucher and J. R. Driscoll, "Automated Abstracting Using Thematic Roles," Proc. of the Third Pan Pacific Computer Conference, Beijing, China, August 15-19, 1989.

I. Syu and J. R. Driscoll, "Portability in Natural Language Query Interfaces via Thematic Roles," Proc. of the Third Pan Pacific Computer Conference, Beijing, China, August 15-19, 1989.

C. E. Couden and J. R. Driscoll, "Generating Responses with Constructive Contributions for Queries of Expository Text Databases," Proc. of the Third Pan Pacific Computer Conference, Beijing, China, August 15-19, 1989.

S. D. Lang, J. R. Driscoll and J. H. Jou, "A Unified Analysis of Batched Searching for Sequential and Tree Structured Files," Transactions on Data Base Systems (TODS), Vol. 14, NO. 4, December 1989.

J. R. Driscoll, W. Lee, I. Syu, T. G. Rackley and R. D. Capetillo, "Easing the Task of Referencing Electronically Stored Troubleshooting Manuals," Proc. of the Conference of the International Association of Knowledge Engineers (IAKE 89), College Park, Maryland, June 26-28, 1989.

L. Bucher and J. R. Driscoll, "An AI Approach to Automated Abstracting," Proc. of the Fourth Rocky Mountain Conference on Artificial Intelligence, Denver, Colorado, June 8-9, 1989.

J. W. Smith and J. R. Driscoll, "An Artificially Intelligent Text Summarization Algorithm," Proc. of the Fourth Rocky Mountain Conference on Artificial Intelligence, Denver, Colorado, June 8-9, 1989.

I. Syu and J. R. Driscoll, "A Portable Natural Language Query Interface," Proc. of AVIGNON 89: Ninth International Workshop on Expert Systems and their Applications," Avignon, France, May 29-June 2, 1989.

J. R. Driscoll, D. A. Rajala, W. H. Shaffer and D. W. Thomas, "An Application of Artificial Intelligence Techniques to Automated Keywording," Proc. of RIAO88 Conference on User-Oriented Content-Based Text and Image Handling, MIT, Cambridge, Massachusetts, March 1988.

J. R. Driscoll, S. D. Lang and L. A. Franklin, "Modeling B-tree Insertion Activity," Information Processing Letters, Vol. 26, No. 1, September 1987.

J. R. Driscoll, S. D. Lang and S. M. Bratman, "Achieving Minimum Height for Block Split Tree Structured Files," Information Systems (GB), Vol. 12, No. 1, 1987.

J. R. Driscoll, H. N. Srinidhi and T. S. Chesser, "A Network Technique to Achieve Program and Data Security with Nominal Communications Overhead," Proc. of the 1986 ACM/IEEE Fall Joint Computer Conference, Dallas, TX, November 1986.

S. D. Lang, J. R. Driscoll and J. H. Jou, "Batch Insertion for Tree Structured File Organizations--Improving Differential Database Representation," Information Systems (GB), Vol. 11, No. 2, 1986.

S. D. Lang, J. R. Driscoll and J. H. Jou, "Improving the Differential File Technique via Batch Operations for Tree Structured File Organizations," Proc. of the Second Annual IEEE Data Engineering Conf., Los Angeles, CA, February 1986.

K. Kinsley and J. R. Driscoll, "A Generalized Method for Maintaining Views," Proc. of the 1984 National Computer Conference (AFIPS press), Las Vegas, Nevada, June 1984.

S. Masters and J. R. Driscoll, "An Evaluation of RQL: A Relational Data Base System for a Low-End Microcomputer Configuration," ACM SIGSMALL Newsletter, May 1983.

S. Masters and J. R. Driscoll, "RQL: A Relational Data Base System for a Low-End Microcomputer Configuration," Proc. of Workshop on Relational DBMS Design/Implementation/Use on Microcomputers, Toulouse, France, February 1983.

K. S. Barley and J. R. Driscoll, "A Survey of Data Base Management Systems for Microcomputers," BYTE Magazine, November 1981.

R. C. Brigham, R. D. Dutton and J. R. Driscoll, "Complexity of a Proposed Data Base Storage Structure," Information Systems (GB), Vol. 6, No. 1, 1981.

V. Gharavi and J. R. Driscoll, "The Implementation of Dynamic Derived Relations," Proc. of the COMPSAC Fourth International Conference, Chicago, Illinois, October 1980.

C. H. Chen, J. R. Driscoll and K. Grammel, "Physical Storage and Physical Navigation Within the Relational DBMS RAQUEL II," Proc. of the Eighth Texas Conference on Computing Systems, Dallas, Texas, November 1979.

. A. Dutton, C. H. Chen and J. R. Driscoll, "A Relational DBMS Conforming to an Architecture Which Incorporates a Physical Storage Language and a Physical Navigation Language," Proc. of the COMPSAC Third International Conference, Chicago, Illinois, November 1979.

`. Kinsley and J. R. Driscoll, "Dynamic Derived Relations Within the RAQUEL II DBMS," Proc. of the ACM 79 National Conference, oit, Michigan, October 1979.

T. Bonar and J. R. Driscoll, "A Very Easy Hierarchical DBMS Implementation," Proc. of the ACM 79 National Conference, Detroit, Michigan, October 1979.

J. R. Driscoll, B. A. Dutton and K. C. Kinsley, "A Relational Storage Scheme Suitable for Derived Views," Proc. of the ACM 78 National Conference, Washington, D.C., December 1978.

J. R. Driscoll, "Two Languages, and their Environment, Founded on Basic DBMS Constructs," Proc. of the Third Jerusalem Conference on Information Technology, Jerusalem, August 1978.

J. R. Driscoll and Y. E. Lien, "A Selective Traversal Algorithm for Binary Search Trees, "Communications of the ACM, U.S.A., June 1978.

J. R. Driscoll, "The Physical Representation of Data Models," Proc. of ICMOD 78, Milan, Italy, June 1978.

Y. E. Lien, C. E. Taylor, J. R. Driscoll and M. L. Reynolds, "Binary Search Tree Complex--Towards the Implementation of Relations," Proc. Int. Conf. on Very Large Data Bases, Framingham, Massachusetts, September 1975.

GRANTS

NASA Kennedy Space Center, "Intelligent Interactive High Speed Data Search System (Year 2, Year 3)," J. Driscoll, Summer Semester 1990 through Fall Semester 1992, $330,000 (approved but award pending).

NASA Kennedy Space Center, "Intelligent Interactive High Speed Data Search System (Phase II)," J. Driscoll, Summer and Fall Semester 1990, $68,361.00 (awarded June 1990).

NASA Kennedy Space Center, "Intelligent Interactive High Speed Data Search System (Phase I)," J. Driscoll, Spring Semester 1990, $82,874 (awarded December 1989).

NASA Kennedy Space Center, "Public Affairs Q&A System," J. Driscoll, Summer Semester 1989, $33,209 (awarded May 1989).

DoD Army Research Institute/PM-Trade, "Conceptual Information Extraction and Summarization for Expository Text," J. Driscoll, ll Semester 1988, $27,321 (awarded August 1988).

Institute for Simulation and Training (IST), "The Application for Statistical Techniques to Improve the JULLS Automated Keywording System," L. Malone and J. Driscoll, Fall Semester 1987, $9,951 (awarded August 1987).

Institute for Simulation and Training (IST), DoD Training and Performance Data Center (TPDC), "Data Base and AI System Technology," J. Driscoll - Salary support for Fall Semester 1987 and Spring Semester 1988, $25,142 (awarded August 1987).

Institute for Simulation and Training (IST), DoD Training and Performance Data Center (TPDC), "Automated Keywording," J. Driscoll - Salary support for Summer Semester 1987, $15,963 (awarded May 1987).

Institute for Simulation and Training (IST), DoD Training and Performance Data Center (TPDC), "Automated Keywording, Collective Training Report, and Functional Description," J. Driscoll - Salary support and publication expenses for Spring Semester 1987, $13,112 (awarded December 1986).

Institute for Simulation and Training (IST), DoD Training and Performance Data Center (TPDC), "Joint Universal Lessons Learned System," J. Driscoll - salary support for Summer Semester 1986, $20,930 (awarded May 1986).

University of Central Florida In-House Grant, "Computer Program and Data Security," J. Driscoll - Salary support for Summer Semester 1985, $5,112 (awarded March 1985).

INTEL Computer Corporation, "Minimization of Logic and Data Traffic in Distributed Networks," J. Driscoll and H. Srinidhi - computer equipment and software, $40,000 (awarded December 1984).

Contract, "INSYSTEC Corporation and the University of Central Florida Purchase Agreement," J. Driscoll - salary support on an hourly basis for "on-campus" and "off-campus" work (awarded November 1981).

NSF LOCI Grant, "Total Microprocessor System Design and Development," R. Guha and J. Driscoll - salary support until September 1983, $46,248 (awarded September 1981).

University of Central Florida In-House Grant, "An Evaluation of Two Methods for Implementing Dynamic Derived Relations," J. Driscoll - salary support for Summer Quarter 1980, $2,340 (awarded January 1980).

University of Central Florida In-House Grant, "A Rational Data Base Management System Supporting Dynamic Derived Relations," J. riscoll - salary support for Summer Quarter 1979, $3,008 (awarded November 1978).

University of Central Florida In-House Grant, "A Data Base Management Systems Supporting Multiple User Views of a Single Data Base," J. Driscoll - salary support for Summer Quarter 1978, $2,780 (awarded November 1977).

CURRENT INTERESTS AND RESEARCH

Data Base Management Systems
Systems Design
Analysis of Algorithms and Data Structures
AI Based Intelligent Data Base Systems
Expert Systems
Hypertext and Hypermedia

DATA BASE SYSTEMS EXPERIENCE

dBASE III Plus
Clipper
FOCUS
IMS
Oracle (on Harris Computer)

ACHIEVEMENTS

Developed a novel Sprinkling System Controller using microprocessor technology. A patent is being applied for.

Developed a relational DBMS for the Apple II microcomputer which was sold nationally as an educational tool via HELLO Software.

Developed several data base systems for the IBM PC which were also distributed to universities for use in data base courses.

NUMBER OF DOCTORAL STUDENTS DIRECTED: 1

NUMBER OF MASTER'S STUDENTS DIRECTED: 20

COMMUNITY SERVICE

Faculty Advisor for the University of Central Florida ACM Student Chapter (September 1976 through June 1980)

Acting Chairman for the Central Florida Chapter of the ACM (October 1978 through April 1980)

RECOGNITIONS

1971	The award of Highest Distinction in the completion of the B.S. Degree by the University of Kansas.
1974	The award of Honors in completion of the requirements of the M.S. Degree by the Department of Computer Science at the University of Kansas.
1979	Listed in Who's Who in Technology.
1982	Awarded second place for the article "A Survey of DBMSs for Microcomputers" in the reader's poll of the November 1981 issue of BYTE Magazine.

PROFESSIONAL ORGANIZATIONS

Association of Computing Machinery (ACM)
IEEE Computer Society
International Association of Knowledge Engineers (IAKE)

OTHER SUPPORT RECEIVED OR PENDING

NASA Kennedy Space Center, "Intelligent Interactive High Speed Data Search System (Year 2, Year 3)," J. Driscoll, Summer Semester 1990 through Fall Semester 1992, $330,000 (approved but award pending).

NASA Kennedy Space Center, "Intelligent Interactive High Speed Data Search System (Phase II)," J. Driscoll, Summer and Fall Semester 1990, $68,361 (awarded June 1990).

EXTERNAL SUPPORT, INDUSTRIAL/FEDERAL COLLABORATION

NASA Kennedy Space Center, "Intelligent Interactive High Speed Data Search System (Year 2, Year 3)," J. Driscoll, Summer Semester 1990 through Fall Semester 1992, $330,000 (approved but award pending).

NASA Kennedy Space Center, "Intelligent Interactive High Speed Data Search System (Phase II)," J. Driscoll, Summer and Fall Semester 1990, $68,361 (awarded June 1990).

NASA Kennedy Space Center, "Intelligent Interactive High Speed Data Search System (Phase I)," J> Driscoll, Spring Semester 1990, $82,874 (awarded December 1989).

NASA Kennedy Space Center, "Public Affairs Q&A System," J. Driscoll, Summer Semester 1989, $33,209 (awarded May 1989).

DoD Army Research Institute/PM-Trade, "Conceptual Information Extraction and Summarization for Expository Text," J. Driscoll, Fall Semester 1988, $27,321 (awarded August 1988).

OTHER SUS INTERACTIONS

No collaboration exists at the present time, but during the proposed planning phase, an effort will be made to identify possible interactions.

TECHNOLOGY TRANSFER

Beyond the usual techniques of publication in conference proceedings and journals, the principal investigator is planning an intelligent text retrieval product demonstration in November 1991 at the Hypertext 91 Conference.

ADDITIONAL SUPPORTING DOCUMENTATION

1. RIAO 91 Intelligent Text and Image Handling - Call for Papers

2. RIAO 91 Intelligent Text and Image Handling - Call for Product Demonstrations

3. DATABASE Journal Article about SPIRIT

stamp

RIAO 91
Center for the Advanced Study of Information
Systems, Inc. (CASIS)
Ms. M.-T. MAURICE
220 East 72nd Street #10F
New York, N.Y. 10021
U.S.A.

Paper submitted (4 copies, 20 pages maximum) must arrive before October 30, 1990:

* CID :
26 bis rue Baffin F-75009 PARIS FRANCE
Tel.: (33) (1) 42 65 64 75 Fax : (33) (1) 45 26 64 45
* North America
Ms M.-T. MAURICE
220 East 72nd Street #10F New York NY 10021 USA
Tel (212) 879 48 19 Fax (212) 685 81 86

Each oral presentation will last 20 minutes followed by 10 minutes of discussion.

Certain papers will be published in a more detailed presentation in the journal: "INFORMATION PROCESS AND MANAGEMENT"

English will be the working language of the conference.

PROGRAM COMMITTEE
Chairman
J. ARSAC
Professor of University of Paris VI

J. ANDRE (F) IRISA Rennes
J.C. BASSANO (F) Univ. d'Orléans
H. BAUER-BERNET (B) Univ. Libre de Bruxelles
A. BOOKSTEIN (USA) Univ. of Chicago
F. BRUNET (SP) Univ. Polytécnica Catalunya
C. CHEN (USA) Simmons College
C. CHRISMENT (F) Univ. de Toulouse
M. CREHANGE (F) Univ. de Nancy
M. DILLON (USA) OCLC
J. DRISCOLL (USA) Univ. of Central Florida
C. FLUHR (F) Univ. ParisXI/INSTN
E. FOX (USA) Virginia Polytech. Inst.
C. FRASSON (CAN) Univ. de Montréal
R. GALLAIRE (F) Bull SA Paris
A. GAMACHE (CAN) Univ. Laval
F. GARCIA-CAMARERO (SP) Univ. Complutense Madrid
P. GOOSSENS (B) Bibliothèque Royale
G. GREFENSTETTE (F) Univ. Tours
R. HANSEN (S) CEB DOXUN
C. JONCKHEERE (NDL) Inst. European Bureau
R. KUHLEN (FRG) Univ. de Constance
A. LELU (F) CNRS/INIST Nancy
B. LUCARELLA (I) Univ. Milan
J. MULLENDERS (B) Alcatel/Bell
S. OHSUGA (JP) Univ. of Tokyo
A.P. PARKES (GB) Univ. of Lancaster
A. PAROCH (F) LIMSI/Paris XI
L. RAU (USA) General Electric
J. ROUAULT (F) Univ. de Grenoble
T. SARACEVIC (USA) Rutgers Univ.
C. SCHWARZ (FRG) Siemens
S. TURNER (F) CNRS/DIST Nancy
B. VICKERY (GB) [illegible]
J.J. VILLANUEVA (SP) U.A.B.
D. WALKER (USA) Bell Comm. Res.
A. ZAMPOLLI (I) Univ. of Pisa

Coordinator in Spain:
C. CARACUBERTA U.A.B.

R.I.A.O 91

CALL FOR PAPERS

INTELLIGENT TEXT AND IMAGE HANDLING

Universitat Autònoma de Barcelona
Barcelona, Spain - April 2-5, 1991
Sponsored by the

European Economic Community
and
Minister of "Industrie et Aménagement du Territoire", France
Minister of "Recherche et Technologie, France
President of the "Generalitat de Catalunya", Spain
Rector of the "Universitat Autònoma de Barcelona", Spain

Organized by the:

CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE (C.I.D.)
&
CENTER FOR ADVANCED STUDY OF INFORMATION SYSTEMS, Inc. (CASIS)

with the participation of the
Centre National de la Recherche Scientifique (CNRS)
France Telecom
Institut National de Recherche en Informatique et Automatique (INRIA)
and the
Federación Española de Sociedades de Archivística, Biblioteconomía y Documentación (FESABID)
Instituto de Información y Documentación en Ciencia y Tecnología (ICYT)
American Federation of Information Processing Societies (AFIPS)
This conference is prepared under the direction of
Professor André LICHNEROWICZ
of the Academy of Sciences of Paris

RIAO - Recherche d'Information assistée par ordinateur
(Computer-aided information research)

GENERAL INTRODUCTION

The purpose of this Conference is to present the state of the art in the storage, retrieval and diffusion of non-structured information found in text, image and sound.

This field is developing rapidly : the entire information technology industry is tending towards the convergence of computing, telecommunications and audiovisual techniques. Conditions continually improve for satisfying users' needs and desires for extensive and convivial access to information.

The previous state of the art in this field, "RIAO 88" held at M.I.T. (Cambridge - U.S.A.) in March 1988, was a resounding success, bringing together members of the international scientific community working in these fields.

"RIAO 91" will take place at the Universitat Autònoma de Barcelona (Catalonia, Spain) from April 2-5, 1991. This conference will present, on one hand, recent scientific research, and on the other, demonstrations of prototypes resulting from this research as well as the most innovative new products appearing on the market.

This call for papers is addressed to researchers from all countries, engaged in academic or industrial research.

CALL FOR PAPERS

GENERAL THEME :

Full-text and heterogeneous media data bases are characterized by the fact that the structure of the information that they contain can rarely be known a priori. Traditional hierarchical and relational database management systems provide inadequate treatment. The absence of homogeneous structure and the great diversity of information in even moderately sized bases make it difficult to foresee which sets of questions will be asked. Information research remains a hard problem, yet computing techniques and technologies seem to provide more power than is being used.

You are invited to submit papers showing how these problems of storage, research and diffusion of non-structured information can be solved.

Particular attention will be given to the following themes:

- techniques for reducing imprecision in localizing information in full text
- data input control and verification
- end user interfaces
- new media

A large number of specific subjects can be treated within this general theme.

SPECIFIC THEMES

A - Linguistic analysis for automatic text treatment
  A1. Automatic indexation
  A2. Automatic abstracts
  A3. Natural language interrogation
  A4. Multilingual interfaces
B - Construction and utilisation of large linguistic knowledge bases (electronic dictionaries, thesaurus, bilingual dictionaries)
C - Confidentiality in information retrieval systems
D - Multilingual interrogation and computer assisted translation
E - Automatic extraction of factual information from full text
F - User interfaces and ergonomy of information research systems
G - Artificial intelligence for user aid and for personalizing systems
H - Intelligent navigational aids and automatic data structuring in hypertexts and hypermedia
I - Neural nets for computer aided information research
J - Data entry systems (OCR, automatic structure recognition, document preparation standards...)
K - New applications:
  K1. Software engineering and information research systems, program and documentation retrieval for re-use, production of intelligent tutorial and documentation systems in software development.
  K2. Automatic image indexation via pattern recognition
  K3. Optical memories (videodisk, CD-ROM, CD-I, optical numerical disks)
  K4. Multimedia systems managing text, sound and images
  K5. Content addressable electronic mail systems
  K6. Voice entry and speech recognition

CONDITIONS

In order to be accepted, the papers must be validated by a prototype or a working model. The authors may be asked to demonstrate their prototype or working models to a member of the program committee. We expect the authors to give a demonstration of their system during the conference, at a time separate from their oral presentation. It is advisable that these demonstrations run on standard material.

APPLICATION FORM

NAME : ________ First ________ Last
TITLE/POSITION : ________
ORGANIZATION: ________
ADDRESS: ________
CITY: ________ STATE: ________ ZIP :
COUNTRY : ________ FAX : ________ Electronic mail :
TELEPHONE: ________
I plan to attend the conference, please send me the program: YES NO
I plan to present a paper: YES NO
Conference theme (circle one): A B C D E F G H I J K
Title of the communication:
Are you willing to give a demonstration of your prototype? YES NO
Equipment needed:
Please mail this form before October 30, 1990

SELECTION CONDITIONS

- Please, return the attached "Application Form" before September 30, 1990 with trade documentation concerning the product to be demonstrated.

- A questionnaire will be sent back to you. This must be returned by October 30, 1990.

- A preliminary selection will be made by the committee of experts based on this information, and the selected applications will then be submitted to a more detailed expertise.

- The final decision of the committee will be released on December 20, 1990.

For more information, please contact:
CID :
36 bis rue Ballu
F-75009 PARIS
FRANCE
Tel.: (33) (1) 42 85 04 75
Fax : (33) (1) 46 26 84 45

RIAO 91

CALL FOR PRODUCT DEMONSTRATIONS

INTELLIGENT

TEXT AND IMAGE HANDLING

Universitat Autònoma de Barcelona
Barcelona, Spain - April 2-5, 1991

Sponsored by the

European Economic Community
and the
Minister of "Industrie et Aménagement du Territoire", France
Minister of "Recherche et Technologie", France
President of the "Generalitat de Catalunya", Spain
Rector of the "Universitat Autònoma de Barcelona", Spain

Organized by the:

CENTRE DE HAUTES ETUDES INTERNATIONALES
D'INFORMATIQUE DOCUMENTAIRE
(C.I.D.)

with the participation of the
Centre National de la Recherche Scientifique (CNRS)
France Telecom
Institut National de Recherche en Informatique et Automatique
(INRIA)
and the
Federación Española de Sociedades de Archivística, Biblioteconomía
y Documentación (FESABID)
Instituto de Información y Documentación en Ciencia y Tecnología
(ICYT)
American Federation of Information Processing Societies (AFIPS)

This conference is prepared under the direction of
Professor André LICHNEROWICZ
of the Academy of Sciences of Paris

RIAO = Recherche d'information assistée par ordinateur
(Computer-aided information retrieval)

GENERAL INTRODUCTION

The purpose of this Conference is to present the state of the art in the storage, retrieval and diffusion of non-structured information found in text, image and sound.

This field is developing rapidly : the entire information technology industry is tending towards the convergence of computing, telecommunications and audiovisual techniques. Conditions continually improve for satisfying users' needs and desires for extensive and convivial access to information.

The previous state of the art in this field, "RIAO 88" held at M.I.T. (Cambridge - U.S.A.) in March 1988, was a resounding success, bringing together members of the international scientific community working in these fields.

"RIAO 91" will take place at the Universitat Autònoma de Barcelona (Catalonia, Spain) from April 2-5, 1991. This conference will present, on one hand, recent scientific research, and on the other, demonstrations of prototypes resulting from this research, as well as the most innovative new products appearing on the market.

The call for papers is addressed to researchers from all countries, engaged in academic or industrial research.

This call for product demonstrations is addressed to European companies or organizations, marketing hardware or software related to the conference themes.

CALL FOR PRODUCT DEMONSTRATIONS

Companies and organizations which market innovative and competitive hardware or software related to the conference themes should exhibit at RIAO 91 for these reasons:

- The products selected for demonstration (no more than thirty) are chosen by a committee of experts on the basis of innovation, and responsiveness to current and future market needs. This selection is itself a guarantee of quality which can be used to the product's advantage.

- A free space will be reserved for the demonstration of the industrial products selected.

- RIAO 88 has shown that this exhibition has been very beneficial, as previously selected companies will testify.

- At RIAO 91, the marketing agent wil be able to contact potential clients directly, gauge the competition, and meet the best specialists in the domain who may be able to help in future development.

- The participants will benefit from the extensive communication effort associated with the exhibition: a catalog, a videocassette, numerous articles in the international press, as well as a deal-making session organized during the conference to bring together marketing, development and research partners.

GENERAL THEMES

The exhibition will highlight hardware and software that facilitate the storage, management, retrieval and diffusion of non-structured information present in text, images, or sound. A list of specific themes follows:

SPECIFIC THEMES

A - Text retrieval systems incorporating:
- A1 . Automatic indexation using linguistic tools
- A2 . Automatic abstracting
- A3 . Natural language interrogation
- A4 . DBMS incorporating full text
- A5 . Multilingual interfaces
- A6 . Speech interfaces
- A7 . Real-time updating
- A8 . Hypertext, hypermedia
- A9 . Query reformulation
- A10. Inter-system gateways

B - Information entry systems: B1 . Optical character recognition
- B2 . Standards (SGML, ODA, ...)
- B3 . Sound, image, and text compression

C - Archiving systems (CD-ROM, CD-V, CD-I, CD-XA...)

D - Help systems
- D1. Lexical data bases
- D2. Conceptual graphs

E - New visualization technologies (HDTV, imagers...)

F - User environment:
- F1. Specialized architectures
- F2. Networks
- F3. Workstations
- F4. Electronic mail
- F5. Computer assisted translation

OTHER DATA REQUIREMENTS

1. **Personnel**

2. **External Support**

Page intentionally left blank

Sate of Florida
High Technology and Industry Council
Applied Research Grants Program

Faculty Profile Form

Personal Information

Name

Please Print

Driscoll
(Last)

James R.
(First) *(Middle Initial)*

Title Associate Professor

Dept/Institute/Center Dept. of Computer Science

College College of Arts & Sciences

University University of Central Florida

Expertise Narrative

Please describe in detail your overall areas of expertise as they pertain to this project

In the last four years, Dr. Driscoll has produced eighteen international conference and journal publications in the areas of efficient file organization, and automatic text processing (automatic keywording, text retrieval, and intelligent text manipulation). For the past three years, he has been continuously funded in these areas by DoD Training Performance Data Center (TPDC), Army Research Institute (ARI)/PM-Trade, and NASA Kennedy Space Center. Total funding from these sources has been $627,000. The work funded by DoD TPDC led to the development of a computer system which mimicked the behavior of an expert human keyworder and assigned keywords to military lessons learned from war games, excercises, and real military confrontations. The system is in use today. The work funded by ARI/PM-Trade involved a demonstration that general world knowledge (surface knowledge) could be used to automate the activity of searching military maintenance manuals in order to match problem descriptions to symptoms in search of corrective procedures. The work funded by NASA KSC concerns efficient retrieval of text information. Dr. Driscoll is a member of the program committee for RIAO '90 International Conference on Intelligent Text and Image Handling.

Sate of Florida
High Technology and Industry Council
Applied Research Grants Program

Industry Profile Form

Company Name

Please Print

John F. Kennedy Space Center, NASA

Address

Kennedy Space Center (City) Florida (State)

32899 (Zip)

Expertise Narrative

Please describe in detail your company's research and development activities

The Advanced Operations Program within NASA is a technology demonstration program which seeks, through its projects, to enhance shuttle operations by the reduction of costs, the increase of efficiency, and the improvement of safety. A majority of the funds expended by the program are to support software projects because it is through the application of artificial intelligence, expert systems, database systems, and other newly developed software techniques that the most significant near-term benefits to shuttle operations are foreseen. At KSC, the projects include Remote Maintenance Monitoring System, Operations Analyst for a Distributed System, Intelligent Launch Decision Support System, Main Propulsion System Pneumatic, Intelligent Computer Aided Trainer, Computer CARE Center, Natural Language Knowledge Acquisition System, Ground Operating Simulation Technique, Intelligent Interactive High Speed Data Search System, and Intelligent Interactive Visual DBMS.

Point of Contact

For additional information concerning this company/business, please contact:

Name

Please Print

Davis (Last)

Tom (First)

(Middle Initial)

Title NASA Technical Officer for Cooperative Agreement NCC10-0003 S-2

Telephone (407) 867-2780 (Area Code)

Section 5.

An Analysis of Natural Language Questions

Index:

1. Non-question information

A. Thematic Roles and their definitions(including alternate names.

1) Verbs

a) ACTION denotes movement or activity that can be seen or heard. ex: Ice drips. ACTION can also be designated affective or causative.

b) PROCESS refers to internal activities (i.e seeing, hearing) or to changes in condition of persons or things that are experiencers or patients. ex: Ice melts. PROCESS can also be designated affective or causative.

c) STATIVE denoting persons or things that are in a particular state or condition. Verbs in Patterns 3, 4, and 5 below are stative most of the time. "However, verbs such as STAY and REMAIN, which may be stative verbs, are commonly intransitive." STATIVE can also be designated ambient, static, or dynamic.

2) Other Words

a) MOVER designates a person or animate creature that performs an action.

b) PATIENT a person or object on whom an action is performed or who receives the effect of an action or process.

c) EXPERIENCER a person or animate being that undergoes or experiences a change.

d) INSTRUMENT subject of an action or process verb OR an object that must be held and manipulated by some person. Something that has a part in bringing about an action or process but is not the instigator. (Can this also include manipulation by a machine?)

e) ENTITY person or thing in a particular state or condition (subject of stative verbs)

f) AGENT the performer of an action that affects another being.

g) COMPLEMENT inanimate or intangible things that come into being as the result of an action or process. They do not receive the effect of an action.

EX/ Build a bookcase. ACTION COMPLEMENT
Dust a bookcase. ACTION PATIENT

h) BENEFICIARY/RECIPIENT person or animate thing that profits or benefits from an action or process. Verbs such as "receive", "accept", "get" have beneficiaries.

i) POSSESSOR someone or something that owns or possesses something.

EX/ John has a parrot. POSSESSOR PROCESS PATIENT

j) PART parts of entities.

EX/ Dogs have paws. ENTITY STATIVE PART

k) ATTRIBUTE the subcases of modifiers. They include: size, age, condition, shape, color, quality(inherent characteristic), condition(not inherent and/or subject to change) see p 50 (VH).

l) EQUIVALENT a noun with a syntactic role as a predicate noun is equivalent to the entity.

EX/ Mary is a Nurse. Tom will become a doctor.
The apple turned to gold.
The girls stayed friends.
He remained a teacher.
John became a firefighter.

m) REASON/CAUSE see p 138-140 (VH)

n) TIME This thematic role appears to include the subroles of FREQUENCY and DURATION.

o) LOCATION The place of the occurence.

p) MANNER this thematic role includes the subrole of INTENSIFIER.

q) CONCESSION this role is pointed to by "although" or "though".

r) EFFECT/CONSEQUENCE

B. Sentence Patterns

1) Simple sentences - simple, active, declarative, positive (VH)
Formats: NP = Nounphrase
Vi = Intransitive verb(doesn't need a NP or adj to complete its meaning)
V = Mostly transitive verbs which allow a passive voice, plus "have", "cost", weigh".
Vl = Linking verb that links an adjective to a NP. Some linking verbs (STATIVE) are: be, taste, feel, smell, look, sound, turn, grow, seem, become, remain, stay, appear
Vbe = A form of the verb "to be".
( ) = Optional part
Adj = Adjective
Adv = Adverbial

a) Pattern 1 : NP + Vi + (Adv)
The adverbials allowed in Pattern 1 are those of location, manner, time, frequency, and duration.
The Pattern 1 particles are: down, up, around, out, in.

b) Pattern 2 : NP1 + V + NP2 + (Adv)
All Pattern 1 adverbials are allowed plus indirect object phrases. The most common transitive verbs that allow both direct and indirect objects are:

ask	buy	give	offer	read	send	teach	throw
bring	find	make	pay	sell	throw	tell	toss

Many Pattern 2 verbs have particles as complements. These are called "two word verbs" or "double verbs" in the literature. The particles are:
off, out, up, on, over, away, down, in

c) Pattern 3 : NP + Vl + Adj + (Adv)
Np is the subject of this sentence. In Pattern 3, you cannot use an adverbial of MANNER with be, remain, or seem. Pattern 3 adverbials are of time, location, duration, and frequency.

d) Pattern 4 : NP1 + Vl + NP1 + (Adv)
NP1 stands for the subject and its EQUIVALENT.
The linking verbs can include particle complements.
EX/ turn into, turn to, change to, change into.

e) Pattern 5 : NP + Vbe + Adv + (Adv)
The first adverbial is obligatory and is most usually location or time.

2) Compound sentences (RH)

Definition: A clause is a group of words containing a nounphrase and a verbphrase that is part of sentence.
Definition: A Main clause is a clause that can stand alone as a sentence.

A Compound sentence contains 2 or more main clauses. They are separated by conjunctions such as "and", "or", "but", "nor", or "so". In some cases, the conjuction is a comma or semicolon.

3) Complex sentences (RH)

Definition: A subordinate clause functions as a dependent clause within a larger construction that is itself a clause or a constituent of one. It cannot stand alone.

A Complex Sentence contains a main clause and one or more subordinate clauses.

2. Questions

A. List of categories

1) Open Questions (RH) or Wh-questions (VH).
The set of possible values for the variable is open or infinite.

ex: Who wrote the letter? -> What person(X) wrote the letter?

Person(X) can be : She or John or anyone.

a) Without Inversion (VH) - Substitution of WH word for variable found as the subject of a sentence or as a modifier of the subject. The Wh- words that can be used in this category are:
Who, what, whose, which.

Ex: Who is an astronaut? -> SUBJECT, AGENT
This question requires a nounphrase answer.

What is in the cargo bay? -> SUBJECT
Whose parts are used to build the shuttle? -> Modifier
This question requires a THEME or Subject modifier response.
Which shuttle flew the longest? -> SUBJECT

b) With Inversion (VH) - When information is contained in an optional role, Theme(Object), or pertains to an adverbial, copula (be verb) or auxiliary verb inversion must occur.
The WH- words that can be used in this category are:
What, when, where, why, how, and who (if theme).

EX: What is the crew doing? -> THEME
When is the crew leaving? -> TIME
Where is the launchpad? -> LOCATION
Why is the shuttle yawing? -> CAUSE/REASON
How can Bill study all night? -> MANNER, ACTION
Who make up the crew? -> BEMOD, THEME
- This question requires a nounphrase response.

c) With DO support (VH)

1) For the auxiliary inversion in simple present or past tense. The WH-words that can be used in this category are:
What, why, how, where, when, whose, which.

EX1 by evolution:

Statement	He broke (something).
DO support	He did break (something).
AuxInversion	Did he break (something/what)?
WH-question	What did he break?

EX2: John broke Mary's stick for the fire by bending it across his knee.

1.	What did John break?	THEME
	What kind of stick did he break?	modifier of THEME or Subject
2.	Why did John break the stick?	REASON/CAUSE
3.	How did John break the stick?	MANNER
	How many sticks did John break?	modifier of
	How much milk did you get?	THEME or AGENT
4.	Where did John break the stick?	LOCATION
5.	When did John break the stick?	TIME
6.	Whose stick did John break?	THEME Modifier.

2) Questions asking about the activity (action) are of the 'what . . . do' form and require a verbphrase or sentence response.

Ex: What did the astronauts do? <- DO support
What will they do now? <-|
What had the support crew done? |- have only aux
What is the astronaut doing? <-| inversion

d) With prepositions (VH) -

EX: info wanted	Question	Statement/answer
to/for+IO	To whom did he send the gift?	He sent the gift to (someone).

This question requires a nounphrase response stating the beneficiary.

location maybe destination?	To what city are they moving?	They are moving to (some city).
time	At what time is the meeting?	The meeting is at (some time).
reason/cause	For what did he buy the item?	He bought the item for (some reason).

These could also have been expressed using Who, Where, When, and Why.

e) Questions requiring a summarization or recounting of events begin with: 'What happened'.

Ex: What happened to the Challenger?
What happened after the Challenger blew up?
What happened at the control center?

f) Echo Questions (RH & VH) NOT TO BE CONSIDERED IN QUERY SYSTEM

Questions where the Wh-word is used for what the individual wants repeated for affirmation or wants to express surprise about.

EX: He did WHAT? The shuttle is leaving WHEN?
The shuttle is HOW tall?

g) Wh-questions requiring Adverbial Responses (VH)

The Wh-words that can be used in this category are: Where, when, why, how long(CONFUSING), how often, how, and how(adj).

```
Ex:  where              -> LOCATION
     when(what time)    -> TIME
     why(what...for)    -> REASON/CAUSE
     how long           -> DURATION
     how often          -> FREQUENCY
     how                -> MANNER
     how (adj)          -> BEMOD, or topic modifier
                             how big (tall,heavy,cold)
```

questions can begin with prepositional phrases except when containing an indirect object: In what city, at what time, to what city. . .

2) Closed Questions (RH)

a) Yes-No (VH) and (RH)

The set of answer values is yes or no. Ex: Is he finished?

b) Or (RH)

The set of alternative responses are in the question.

Ex: Is the shuttle in CA or FL?

B. Categories for each Wh-word

1) Who 1a, 1b, 1d

2) What 1a, 1b, 1c, 1e

3) When 1b, 1c, 1d, 1g

4) Why 1b, 1c, 1d, 1g

5) Where 1b, 1c, 1d, 1g

6) How 1b, 1c, 1g

7) Whose 1a

8) Which 1a, 1c

9) Whom 1d

C. Categorization of 35 questions (List attached)

	Question	Category
1)	What is the maximum cargo weight the shuttle can carry?	1a
2)	How far can the shuttle transport cargo from the earth's surface?	1g
3)	What has happened to the Enterprise?	1e
4)	How many years of education are required for astronaut candidates?	1c1
5)	What is the total weight of the shuttle?	1a
6)	How thick is the window of the shuttle?	1g
7)	How many gallons of liquid hydrogen fuel can the storage tank hold?	1c1
8)	What type of liquid fuel is used on the shuttle?	1c1
9)	What is the descent rate of the shuttle during landing?	1a
10)	How long is the machanical arm used for payload deployment?	1g
11)	What are the dimensions of the cargo area in the shuttle?	1a
12)	How is waste disposed of?	1b
13)	Have there been astronauts picked from minority groups?	2a
14)	What is the total number of times that the shuttle has been launched?	1a,b
15)	What type of food do astronauts eat during a shuttle mission?	1c1
16)	What is the orbiter's velocity while in orbit?	1a,b
17)	What is the maximum acceleration of the shuttle during launch?	1a,b
18)	What is the maximum touchdown glide speed of the shuttle?	1a,b
19)	How many pounds of thrust do the SRB booster rockets generate during liftoff?	1c1
20)	What is the maximum fluid fuel flow rate during launch?	1a,b
21)	How fast does the crawler or transporter travel?	1g
22)	At what altitude and speed must the pilot throttle back during ascent?	1b
23)	a) How many general purpose computers are on board the shuttle?	1c1
	b) What functions do they serve?	1c2
24)	What is the new design of general purpose computers like on board the shuttle?	1b

25) What is the total number of tiles that cover the orbiter for thermal protection during reentry? 1a,b

26) When did the first space shuttle launch occur? 1g

27) How long is the runway at a shuttle landing facility? 1g

28) What type of glass is used for the windows to withstand the pressure of flight 1c1

29) What is the total amount of RAM available in the shuttle's general purpose flight computer? 1a,b

30) What material are the heat shield tiles composed of? 1b

31) What type of computer guidance and navigation systems does the shuttle use during reentry and landing? 1c1

32) What is the maximum power available to the payload area? 1a,b

33) Are there emergency escape prodedures to jettison the crew members out of the shuttle? 2a

34) Where do the crew members sleep on the shuttle? 1g

35) What is the color of the external tank? 1b,c

D. Question Type Count

1a	1b	1a or b	1b or c	1c1	1c2	1e	1g	2a
4	4	8	1	8	1	1	7	2

E. Question Types in TRQS (JD)

		Category
1)	Query what can be asked Cue words: Subjects or Topics. Answer: All Subjects and Objects in target database.	1b
2)	Query the existence of a fact Cue words are: Who, Which, Whom, and What. Answer: Complete sentence.	1a,b,c,d,e
3)	Query description in detail of an event. Cue words are: Who, Which, Whom, and What. Answer: Complete sentence.	1a,b,c,d,e
4)	Query LOCATION of an event Cue word: Where. Answer: Location Thematic Role.	1b,c,d,g
5)	Query SOURCE of an event Cue word: Where. Answer: Source Thematic Role.	1b,c,d,g
6)	Query the DESTINATION of an event No cue word or answer set that will return the Destination Thematic Role.	
7)	Query the TIME of an event Cue word: When. Answer: Time Thematic Role.	1b,c,d,g
8)	Query the person or thing that causes an event Cue words are: Who, Which, Whom, and What. Answer: Complete Sentence.	1a,b,c,d,e
9)	Query the INSTRUMENT or CONVEYANCE of an event Cue words are: Who, Which, Whom, and What. Answer: Complete Sentence.	1a,b,c,d,e
10)	Query the BENEFICIARY of an event Cue words: Benefited From. Answer: Beneficiary Thematic Role.	1a,b,d
11)	Query the PURPOSE of an event Cue word: Why. Answer: Purpose Thematic Role.	1b,c,d,g

12) Query the DURATION of an event 1g
Cue words: How Long.
Answer: Duration Thematic Role.

13) Query the CO-AGENT of an event 1a,b,c,d,e
Cue words are: Who, Which, Whom, and What.
Answer: Complete Sentence.

14) Query the event which happens to the AGENT or OBJECT 1a,b,c,d,e
Cue words are: Who, Which, Whom, and What.
Answer: Complete Sentence.

Bibliography:

(JD) .. Driscoll, James R. A Portable Natural Language Query Interface for Databases Containing Large Amounts of Textual Data. Manuscript submitted to 16th Very Large Data Base Conference.

(VH) .. Heidinger, Virginia T. Analyzing Syntax & Semantics. Washington, D.C.: Gallaudet College Press, 1984.

(RH) .. Huddleston, Rodney. English grammar: an outline. Cambridge, GB: Cambridge University Press, 1988.

Section 6.

Manuscript about the SPIRIT text retrieval system

from August 1990 DATABASE Magazine

Page intentionally left blank

NEW DIRECTIONS FOR MICROCOMPUTER-BASED HYPERTEXT SYSTEMS

by Kenneth Ray
and
James R. Driscoll

In the second half of the twentieth century major advances have been made in several fields relevant to information management. Among these are the continued development of the microcomputer, high-level programming languages, and sophisticated methods of textual analysis. It is estimated that as of 1985 there were 1.7 billion documents online worldwide [1], and although powerful personal workstations are increasingly common on the desks of professionals, it is a sad reality that even with access to large, continually updated databases of relevant data, most of these professionals do not have the software tools required to effectively locate and organize the available information.

At present, information retrieval (IR) activities are generally conducted using the interfaces provided by commercial databases, under the control of professional search intermediaries or end-users who possess extensive training. These database systems typically reside on mainframe computers and lack the user-friendly interfaces which personal computer users have come to expect. It seems a reasonable goal to build effective, easy to use IR systems which allow workers to focus directly on their problems and tasks, without resorting to intervening technicians or programming languages. This article describes an advanced hypertext environment which couples a probabilistic and linguistic approach to information retrieval with the intuitive, easily browsable document representation characteristic of conventional hypertext systems.

(Editor's Note: Readers needing a refresher on conventional hypertext systems should read Carl Franklin's article, "Hypertext Defined and Applied," in the May 1989 issue of ONLINE *13, No. 3, pp. 37-49. —PH)*

COMMERCIAL INFORMATION RETRIEVAL SYSTEMS

One of the reasons people do not find what they are looking for in books or textual databases is because the terms they use to describe the things they want are not the terms indexed by the system. In a commercial IR system, based on a Boolean model, this failing can usually be attributed to imprecise or incomplete search terms. It is well known that Boolean queries are not only difficult to construct, but require a trade off between recall (the proportion of appropriate references found) against precision (the proportion of references which are relevant). In practice, a compromise is often obtained by formulating a query which is neither too broad nor too narrow, resulting in a large set of candidate documents, only a few of which will be relevant to the searcher's needs. In general, it is not a good idea to impose Boolean search techniques on a microcomputer user who has no experience in the trade offs required and who may be unwilling or unable to formulate complex queries.

Fortunately, commercial IR systems are now available which have the ability to process natural language queries through probabilistic and linguistic analysis. The information in the textual database of such a system is represented as a collection of index terms which can be matched against a processed query. Those documents whose representations most closely match the query are returned in the form of a ranked list of document classes. This approach has several major benefits. For one, users do not require extensive training as to the system's use, thus freeing them to focus on the task which motivated them to use an IR system in the first place.

These systems employ a hypertext document representation which allows users to browse through the documents gathered in response to a query. The user may choose to read the documents

in a serial fashion, as they were written by the author, or to leap from relevant passage to relevant passage between documents, or to formulate a query from any page of the current document being browsed. When implemented through the intuitive interface typical of conventional hypertext systems, this last ability becomes the basis for a relevance feedback loop. Through successive iterations of query reformulation, the user can "home in" on the collection of documents which best suit his or her interest.

LIMITATIONS OF CONVENTIONAL HYPERTEXT SYSTEMS

The nodes of information which make up a conventional hypertext document are traditionally linked by hand. This is a time consuming task which requires much thought. For applications which use large, dynamic collections of online documentation, it is impractical to attempt the construction and continuous adjustment of a conventional hypertext representation. In contrast, the set of ranked documents gathered in response to a user query on hypertext-based IR systems represents a hypertext structure which requires no maintenance.

Even when a conventional hypertext document has been well designed, it is possible to get lost while browsing through unfamiliar information. While a good author in a paper document will guide the reader through a progression of relevant points, hypertext gives readers the power to guide themselves. The double-edged sword of freedom of movement is both the beauty and the pitfall of conventional hypertext systems. Hypertext-based IR systems solve this problem through the automatic generation of user-defined links, represented by the keywords common to a set of documents. Because the design of this hypertext structure is instantaneously responsive to the will of the framer of a query, the question of where you are in someone elses hypertext network never arises.

ADVANCED HYPERTEXT SYSTEMS

We define an advanced hypertext system as an information retrieval system which processes natural language queries, automatically generates cross-referential hypertext links, includes a relevance feedback system, and maintains the easily browsable document representation characteristic of conventional hypertext systems. We will now discuss a commercial IR product called SPIRIT which represents the basis of such a system.

The SPIRIT system (Syntactic and Probabilistic Indexation and Retrieval of Text) is available through SYSTEX, a French company founded in 1979 by a group of researchers in order to make industrial products in the fields of computational linguistics and artificial intelligence, applied to information retrieval. The SPIRIT system is multilingual with versions in English, French and Arabic. A microcomputer-based, English version of SPIRIT which runs under OS/2 or UNIX should be available by the time you receive this issue of *DATABASE*.

SPIRIT processes natural language queries by computing the semantic proximity between a query and the contents of an indexed textual database, using both statistical and linguistic analysis. SPIRIT does not try to understand a query or the database text; instead, SPIRIT examines each word in a query and database text considering word frequency, word position, and linguistic word forms (e.g., part of speech). The system establishes an ordered list of candidate documents based on their semantic proximity to the query. The linguistic processing of queries includes spelling correction, idiomatic expression recognition, synonym recognition (e.g., "lorry" and "truck" are considered identical), and the resolution of grammatical ambiguities (e.g., "can" as auxiliary verb is distinguishable from its use as a noun). One well-formed natural language query over a selected database can obtain the same result as a complete Boolean search strategy of many Boolean queries. Because natural language is used to express queries, professional search intermediaries are not needed to translate user queries into a form compatible with the indexed terms of a database.

One well-formed natural language query over a selected database can obtain the same result as a complete Boolean search strategy of many Boolean queries.

SOFTWARE PROFILE

SPIRIT

SYSTEX Company
FERME DU MOULON
91190 GIF SUR YVETTE
FRANCE
Telephone: 33 (1) 69 85 33 38
Fax: 33 (1) 60 19 13 12
From the USA it is best to contact SYSTEX via fax.

Cost: from 55,000 FF (single user) to 150,000 FF (multiuser) to 500,000 FF (mainframe)

Hardware/Software Requirements (any of the following systems):

- IBM mainframe using VM or MVS operating system.
- VAX computer using VMS.
- SUN or other workstations using UNIX.
- IBM AT class microcomputers using OS/2 or UNIX.
- Macintosh (available in late 1990).

The version of SPIRIT (2.1) with which the authors are familiar is installed on an IBM 4381 mainframe at the University of Central Florida, and although it does not incorporate synonym recognition, we have been quite impressed with the system's performance in rapidly retrieving relevant text. However, because this version is installed on a mainframe, accessed through character-based terminals, the system screens are not as pleasant as those normally seen on a personal computer.

The indexing of documents in a database is the most time consuming task of using the SPIRIT system. All documents must first be placed in one file (ASCII for microcomputers, EBCDIC for IBM mainframes), with markers separating each document. For example, in demonstrating the SPIRIT system to NASA, Kennedy Space Center, the one thousand page *NASA National Space Transportation System Reference* (better known as the shuttle manual) was marked by considering each paragraph of the manual as a document. This resulted in a database of 4902 documents which required 3.5 hours to index [2]. Of course, a database is only indexed after new material is entered.

FIGURE 1
INITIAL MENU

```
SPIRIT                6.7 R 0.0              - VM / CMS       MENU 0

                 1. SAVING/STORING DOCUMENTS
                 2. CREATE A DATABASE
                 3. UPDATE A DATABASE
                 4. DELETE SOME DOCUMENTS
                 5. DETECTION OF ERRORS

                 S. QUERY THE DATABASE

                 G. MANAGEMENT UTILITIES
                 P. REDEFINE DISK PARAMETERS
                 V. EXECUTE A VM COMMAND
                 F. END

        BY DEFAULT (ENTER) => QUERY THE DATABASE
```

FIGURE 2
MAIN MENU

```
                 ********************************
                 *                              *
                 *           SPIRIT             *
                 *           SYSTEM             *
                 *            R 2.1             *
                 *                              *
                 ********************************

                     PRINCIPAL   MENU
                    __________________

MENU: (QUERY,AFQUERY,CONTQ,BOOL,DOCQ,ANSWER,DOC,BASE,STOP,PRINT,G,?):
```

QUERYING THE SHUTTLE DATABASE USING SPIRIT

SPIRIT is menu-driven and, once set up, is very easy to use. The initial system screen is reproduced in Figure 1. The first five options on the menu provide various database development and maintenance functions. The **S** option of the initial menu allows you to query SPIRIT on the information contained in a database. With the exception of **F**, which terminates the program, the remaining options provide environment specific utilities.

After choosing **S** from the initial system menu, the next SPIRIT screen lists the various databases within the system, the date they were last updated, and prompts you to enter the name of the database you wish to query. Entering a database name causes the main menu to be presented (Figure 2).

The system commands are as follows:

- **QUERY** is used to form a natural language query on the textual fields of a database.
- **AFQUERY** is used to form a natural language query on all fields of the database, not just the textual fields.
- **CONTQ** is used to continue a query, make more than one query simultaneously, or to alter a query.
- **BOOL** is used for Boolean queries.
- **DOCQ** allows all the keywords within a chosen document to be used to query the database.
- **ANSWER** shows the list of documents which satisfied the last query.
- **DOC** is used to choose a document to browse.
- **BASE** is used to change databases.
- **STOP** ends the session and returns to the principal menu.
- **PRINT** is used to print documents.
- **G** (rid queries) allows queries on multiple fields.
- **HISTO** displays a list of the previous questions in a session.

We choose the **QUERY** option from the main menu and enter the question **What is the maximum cargo weight the shuttle can carry?**. After we do this, the screen appears as shown in Figure 3. All questions to the system must end with a question mark, and if SPIRIT detects a word within a query which is misspelled, you will be prompted to re-enter the word. The empty words shown in Figure 3 will be discarded while the keywords will be used to formulate a query.

Whenever *** appears on screen it signals the user to press the return key to continue. A few seconds after pressing the <return> key, a ranked list of document classes appears, as shown in Figure 4. SPIRIT ranks the classes by the number of keywords matched, the proximity of the keywords to each other, and other statistical and linguistic factors. Keywords in close proximity are separated with a hyphen (-). Notice that SPIRIT considers the two keywords in CLASS 2 more relevant to our query than the four

FIGURE 3
NATURAL LANGUAGE QUERY

NATURAL LANGUAGE QUERY ON THE SHUTTLE BASE

<1>: What is the maximum cargo weight the shuttle can carry?

EMPTY WORDS : **what, is, the, the, can.**
KEY WORDS : **maximum, cargo, weight, shuttle, carry.**

***

FIGURE 4
DOCUMENT CLASSES

CLASSES	NB DOCS	KEY-WORDS
1	1	maximum-cargo-weight.
2	1	shuttle-carry.
3	2	cargo-weight, shuttle.
4	2	cargo, weight, shuttle, carry.
5	3	cargo-weight.
6	1	cargo, weight, carry.
7	2	cargo, weight, shuttle.
8	6	cargo, shuttle, carry.
9	1	weight, shuttle, carry.
10	1	maximum, weight.
11	4	cargo, weight.
12	4	cargo, carry.
13	6	maximum, weight.
14	2	maximum, carry.
15	20	cargo, shuttle.
16	9	weight, shuttle.
17	13	shuttle, carry.
BOTTOM OF LIST		

LIST OF CLASSES TO BE DISPLAYED (?) : 1

FIGURE 5
DOCUMENT DISPLAY

DOC 3976 BASE : doc 3976 **NCP:0/CPI:1/NBI:1** **+18 1K/1K**
IDENTIFIER. : doc 3976
TEXT....... :

Cargo weight is defined as the payload control **weight** plus the **weight** of the attached hardware used to secure the payload to the orbiter. Allowable **cargo weight** is determined by altitude and orbital inclination. For example, on a standard inclination of 28.45 degrees, **maximum cargo weight** capability in a circular orbit at an altitude of 100 nautical miles is about 55,000 lb. This capacity decreases with altitude and falls to about 40,000 lb. in a 300-mile circular orbit. At the higher inclination of 57 degrees (also a standard inclination), **cargo weight** capability is 40,000 lb. in a 100-mile circular orbit. This decreases to slightly over 20,000 lb. in a 320-mile orbit. These **weights** are those for a nominal ascent for what is described as a "simple, short duration, satellite deploy mission."
BOTTOM OF DOCUMENT

INFORMATIONAL PAGE 1/1
WHAT DO YOU WANT TO DISPLAY ?
> OR RETURN, <, >>, <<, DOC, END, DDQ, (?) :

FIGURE 6
DOCUMENT DISPLAY

DOC 3976 BASE : doc 3976 **NCP:0/CPI:1/NBI:1** **+18 1K/1K**
IDENTIFIER. : doc 3974
TEXT............ :

For Vandenburg Air Force Base western test range **satellite deploy missions**, using OV-103 or OV-104, the **cargo-lift weight capability** is 29,600 pounds for a **98-degree** launch **inclination** and a **110-nautical-mile (126-statute-mile)** polar **orbit**. Again, an increase in **altitude** costs approximately 100 pounds per **nautical mile.** NASA also assumes that the advanced solid rocket motor will replace the filament-wound solid rocket motor case previously used for test range assessments.
BOTTOM OF DOCUMENT

YOU ARE IN DDQ
WHAT DO YOU WANT TO DISPLAY ?
> OR RETURN, <, >>, <<, DOC, END, DDQ, (?) :

keywords in CLASS 4, and that the combination **shuttle-carry** is more relevant than **cargo-weight**.

One selects the document classes to be browsed by entering a number or list of numbers separated by commas. In our case, the document which contains the keywords **maximum-cargo-weight** is considered to be the most relevant, so we simply enter a **1**. SPIRIT responds as shown in Figure 5.

All keywords are highlighted within the page currently being browsed. The prompt shown below the document includes options for further browsing through the classes of documents in Figure 4, or within the pages of the current document (page up, page down, first page, last page). In our case, we quickly note that this page answers our question and further browsing is not necessary. By pressing **E** for end, we will be returned to the main menu.

BROWSING A DATABASE

On the other hand, if we decide to browse through the entire collection of documents, the repeated pressing of the <return> key will page forward through the current document class, retrieve the next ranked class, and so on until all classes requested have been viewed. If at any time during the browsing of a document class, we encounter a page of information which interests us, but is not directly related to the material our query has gathered, we can employ the **DDQ** option. This option allows the reformulation of a query based on all keywords contained in the current page being viewed.

For example, after reading the paragraph retrieved in Figure 5, suppose we become interested in the effect of altitude and inclination on cargo weight capacity and desire more information relevant to this paragraph. By selecting the **DDQ** option, SPIRIT automatically links us to the most relevant paragraph in the shuttle manual. The result is the screen displayed in Figure 6. Notice that this paragraph provides additional information on the topics which interest us. This represents the most relevant document in a new collection of ranked documents, which can be browsed in the same manner as the previous collection.

▶

CONCLUSION

In a society approaching the state of information overload, it is becoming increasingly apparent that our present methods of retrieving and organizing textual material are too primitive to keep up with the world-wide pace of technological development. The authors believe that the development of hypertext IR systems is an important step in the evolution of information management tools, and that the SPIRIT system embodies the basic requirements of what we define as an advanced hypertext system.

REFERENCES

[1] Williams, M. W. *Proceedings of the Eighth National Online Meeting*, New York, NY. (1987).

[2] Clark, M. A. "A Prototype Text Retrieval System for Answering NASA Kennedy Space Center Media Questions." Research project report submitted to the Department of Computer Science, University of Central Florida, Orlando, FL. (December 1989).

THE AUTHORS

JAMES R. DRISCOLL is an Associate Professor of Computer Science at the University of Central Florida. He has a B.S. degree in electrical engineering and a M.S. and Ph.D. degrees in computer science, all from the University of Kansas. He has published eighteen manuscripts in the last four years in regard to efficient file organization and automatic text processing (automatic keywording, text retrieval, and intelligent text manipulation).

For the past three years he has been continuously funded in these areas of research by the Department of Defense, the Army Research Institute, and NASA. He is a member of the Association for Computing Machinery and the IEEE Computer Society.

KENNETH RAY is a software engineer for Martin Marietta Information Systems in Orlando, Florida.

Communications to the authors should be addressed to James R. Driscoll, Department of Computer Science, University of Central Florida, Orlando, FL 32816; 407/275-2341.